TECHNISCHE UNIVERSITÄT DRESDEN

Bereich Mathematik und Naturwissenschaften Fakultät Psychologie

Bayesian cognitive modeling of the balancing between goal-directed and habitual behavior

Sarah Schwöbel

Dissertation

Supervisors
Prof. Dr. Stefan Kiebel
Prof. Dr. Michael Smolka

Bayesian cognitive modeling of the balancing between goal-directed and habitual behavior

DISSERTATION

zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

vorgelegt

dem Bereich Mathematik und Naturwissenschaften
der Technischen Universität Dresden

von

M.Sc. Sarah Schwöbel

geboren am 26.09.1988 in Heidelberg

eingereicht am 16.06.2020

Gutachter:
Prof. Dr. Stefan Kiebel
Prof. Dr. Tanja Endrass

Die Dissertation wurde in der Zeit von November 2016 bis Juni 2020 im Institut für Allgemeine Psychologie, Biopsychologie and Methoden der Psychologie in der Professur für Neuroimaging unter der Betreuung von Prof. Dr. Stefan Kiebel und Prof. Dr. Michael Smolka angefertigt

Danksagung

Ich danke zuallererst meiner Familie, insbesondere meiner Mutter, meinem Vater und meiner Stiefmutter, ohne deren Unterstützung ich weder einen Studienabschluss noch eine Promotion hätte erreichen können. Ich danke meinen Freunden Nana, Fabi und Arne, ohne deren Hilfe ich diese Promotionsstelle nicht hätte antreten können. Ich danke meinen Kollegen am Lehrstuhl, insbesondere Sascha und Cassandra, für ihre Unterstützung und die inspirierenden Gespräche. Ich danke Dimitrije Marković für seinen wertvollen Input, von dem ich die meisten meiner Modelingfähigkeiten gelernt habe. Ich danke außerdem Michael Smolka für seine Unterstützung und die erhellenden Gespräche. Ein besonderes Dankeschön geht an Stefan Kiebel, für seine wunderbare Betreuung und seine Unterstützung insbesondere in den schwierigen und stressigen Zeiten. Zum Schluss gehört mein größter Dank meinem Partner Daniel, der ohne Beschwerden alle Herausforderungen mitgetragen hat, mir immer den Rücken gestärkt hat, und mich voll und ganz unterstützt hat. Die Wissenschaft ist ein Teamsport und ohne euch wäre ich nur schwer so weit gekommen, vielen Dank an euch alle und all diejenigen, die ich hier nicht explizit genannt habe.

List of publications used in this thesis

Chapter 2 is based on the publication

Schwöbel, S., Kiebel, S. J., & Marković, D. (2018). Active inference, belief propagation, and the bethe approximation. Neural computation, 30(9), 2530-2567.

The study concept was developed by SS under the supervision of SJK and DM. SS developed the models and methods, and analyzed and interpreted the results. SS and DM drafted the manuscript, and SJK provided critical revisions.

This publication is not currently used in any other dissertation, nor is it intended to be used in any future dissertations.

Chapter 3 is based on the preprint

Schwöbel, S., Marković, D., Smolka, M. N., & Kiebel, S. J. (2019). Balancing control: a Bayesian interpretation of habitual and goal-directed behavior. bioRxiv, 836106. (submitted to Journal of Mathematical Psychology)

The study concept was developed by SS under the supervision of SJK, DM, and MNS. SS developed the models and methods, and analyzed and interpreted the results. SS drafted the manuscript, SJK, DM, and MNS provided critical revisions.

This publication is not currently used in any other dissertation, nor is it intended to be used in any future dissertations.

TECHNISCHE UNIVERSITÄT DRESDEN

Bereich Mathematik und Naturwissenschaften Fakultät Psychologie

Abstract

This thesis proposes a novel way to describe habit learning and the resulting balancing of goal-directed and habitual behavior using cognitive computational modeling. This approach builds on experimental evidence that habits may be understood as context-dependent automated sequences of behavior embedded in a hierarchical model. These assumptions were implemented in a Bayesian model, where goal-directed action sequences are encoded using a Markov decision process, and habits are interpreted to arise from a Bayesian prior over such sequences. Simulations show that this modeling approach yields key properties of habit learning, such as increased habit strength with increased training duration. This novel mechanistic description may lead to an improved understanding of habit learning mechanisms and individual learning trajectories, which may have implications for mental disorders which are believed to be accompanied by a maladapted balance between goal-directed an habitual control.

Contents

List of Figures

List of Tables

1 Introduction

Behavior in animals and humans is not only based on the current state of the environment and immediate outcomes of actions. Instead, living agents have developed the remarkable ability to inhibit actions that would currently lead to favorable outcomes in lieu of actions that will yield an even greater outcome in the future. A squirrel collecting and hiding nuts for winter is just as much an impressive example as humans staying at home to study for a degree instead of going out with their friends. To be able to execute such behavioral control, living agents must predict the consequences of their actions over multiple time scales and use this representation of the future to select actions.

This behavior, which is based on internal motivation and future goals instead of being based on current stimuli, impulses, and habits, is often referred to as volitional behavior (Haggard, 2019). Volition itself is viewed to be comprised of a set of executive function or cognitive control processes which guide cognitive processes to adhere to super-ordinate, long-term goals. As a consequence, an agent is enabled to choose behavior reaching goals, even in the presence of conflict from impulses and habits (Goschke, 2014; "The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis", 2000). The exact processing mode of a cognitive process can be described by meta-control parameters, which define the modus operandi of the process at hand and the respective large-scale brain network. Neurobiologically, meta-control parameters have been linked to neuromodulatory systems and a dysfunctional setting of these parameters has been suggested in several mental disorders, such as obsessive compulsive disorder (OCD), addiction, depression, and anorexia nervosa (Goschke, 2014).

Importantly, how meta-control parameters should be configured for healthy and optimal processing is not a trivial question. In such a meta-control problem, an agent has to strike a balance between, for example, an automatic and habitual tendency to execute behavior which was successful in the past, and explicitly evaluating behavior with respect to its consequences in a goal-directed manner. Carefully evaluating actions and planning ahead means an agent can adaptively react to its environment but uses computational resources and time, meaning it carries opportunity costs. Conversely, choosing automatic behavior means faster action and frees mental processing resources for potentially more pressing demands, but behavior may not be adapted to a changing environment. At any time, an agent has to decide whether it can rely on its resource and time efficient habits, or whether it should rather use more costly and slower explicit forward planning.

Consequently, a maladapted balancing between habitual and goal-directed behavior may

be implicated in several mental disorders (Goschke, 2014), and can manifest as an increased learning rate of habits, or an excessive reliance on either habitual or goal-directed behavioral control. OCD for example has been described to be accompanied by an over-reliance on the habit system (Gillan et al., 2011). Another striking example is addiction, where addictive behavior is thought to result from a shift from goal-directed towards habitual control (Volkow & Morales, 2015; Everitt & Robbins, 2005, 2016). Therefore, research into the balancing of habitual and goal-directed behavior may have a sizeable impact on the understanding of how these disorders emerge and how they may be treated.

It has been recently argued that neuroscience cannot solely rely on the bottom-up approach of inferring brain (mal-)function from neuronal architecture, as the analysis and interpretation of neurophysiological data can become rather complex. Instead, it it has been proposed that the neuronal architecture should be viewed to follow function, so that it is imperative to understand behavior and the respective computational necessities in order to be able to understand neuronal architecture (Krakauer, Ghazanfar, Gomez-Marin, MacIver, & Poeppel, 2017; Gomez-Marin, Paton, Kampff, Costa, & Mainen, 2014; Cooper & Peebles, 2015; B. W. Balleine, 2019). By following this approach to gain a more mechanistic and theoretical understanding of habit learning and balancing of control modes, this thesis proposes a novel, hierarchical Bayesian habit learning model, describing habit learning on sequences of actions, which captures key characteristics of habit learning known from the animal literature.

To lay the ground for this new model of habit learning, Chapter 1 will first introduce habit learning in more detail and outlines the behavioral experiments in animals and humans which probe the characteristics of goal-directed and habitual behavior, from which an operational definition of habitual behavior has been derived. This is followed by a short overview over what is known about the neural underpinnings of habitual and goal-directed behavior. Subsequently, models of habit learning will be discussed: Starting with Markov decision processes as a model for goal-directed behavior, followed by an introduction to important advances in modeling of habit learning and the resulting balancing of control. Then approximate Bayesian methods will be discussed which will be used for the model proposed in this thesis. Lastly follows a short overview of the proposed model and which findings from the animal literature it should be able to replicate. Chapter 2 proposes a novel approach to improve sequential inference in Bayesian Markov decision processes based on the belief propagation algorithm. Chapter 3 will present the habit learning model in more detail and show, using simulations, that the model can indeed replicate key behavioral findings from the literature. Finally, Chapter 4 will summarize the findings and integrate them into the larger context of the literature, and discuss implications and limitations should the model predictions hold.

1.1 Operational definition of habitual and goal-directed behavior

The first description of how agents learn and select actions based on reinforcement was Thorndike's "Law of effect" (Thorndike, 1898; Yin & Knowlton, 2006). Translated into current psychological concepts, this kind of learning is nowadays described as operant conditioning or instrumental learning (Mazur, 2015; Gluck, Mercado, & Myers, 2016). Here, it is assumed that an agent, upon encountering a new environment, learns about its structure and what rewards may be achieved and uses this knowledge to navigate to the rewarding goal states in a goal-directed manner. Such goal-directed behavior is typically described as being voluntary

and is operationally defined as being based on learned action-outcome contingencies, where action evaluation and choices are based on the expected reward of an action (Dickinson & Balleine, 1994; Dolan & Dayan, 2013). In the case of a more complex environment, the agent may have to deliberately plan several steps into the future, and use its knowledge about the structure of the environment and action-outcome contingencies to search through a decision tree and evaluate which sequence of actions will lead to a goal (Keeney & Raiffa, 1993). As this process is akin to a mental simulation of the future, it is inherently prospective and future-oriented (Dolan & Dayan, 2013; Yin & Knowlton, 2006; Adams, 1982). The main advantage of such a deliberate planning process is that it is based on the structure of the environment and can therewith be flexibly adapted to changes in the environment, meaning that an agent employing goal-directed behavior is able to flexibly adapt to new action-outcome contingencies (Yin & Knowlton, 2006; Dickinson, Nicholas, & Adams, 1983; Dickinson, 1985). Nonetheless, this comes with a disadvantage: The search through the decision tree can be arbitrarily complex for large state spaces and long planning horizons so that this kind of behavioral evaluation is slow and costly to use (Dayan, 2009; Dolan & Dayan, 2013).

Once action-outcome contingencies have been sufficiently learned for an environment, and an agent has determined which behavior is rewarding, it is more resource efficient to switch to habitual behavioral control (Yin & Knowlton, 2006; Dayan, 2009). Habitual behavior is typically operationally defined as automatic stimulus-response associations, which are based on past behavior and experiences (Dickinson et al., 1983; Graybiel, 2008; Dolan & Dayan, 2013; Smith & Graybiel, 2016). It is inherently retrospective and promotes a repetition of behavior which has been successful in the past (Dayan, 2009; Dolan & Dayan, 2013). With increased training duration in a specific environment, habits become more and more pronounced until the agent is over-trained and habits completely dominate the action selection process (Yin & Knowlton, 2006; Colwill & Rescorla, 1988; Adams, 1982),see Figure 1.1**a**. This comes with the advantage that habits, due to their automatic nature, are fast and resource efficient (Dayan, 2009; Dolan & Dayan, 2013), but on the other hand, behavior may be rendered inflexible to changes in the action-outcome associations and reward values of the environment, especially in over-trained agents (Yin & Knowlton, 2006; Adams, 1982). These operational definitions have been derived from animal experiments which are outlined below.

Animal experiments

The way in which animal experiments usually probe for goal-directed and habitual behavior goes back to Skinner and his operant conditioning chamber (Skinner, 1948; Yin & Knowlton, 2006). This so-called Skinner box provides a confined environment in which animals can be isolated from the rest of the laboratory setting so that stimuli can be controlled by the experimenter. Animals usually have one or two simple and repeatable actions, like lever presses, from which they can choose, see Figure 1.1b, further reducing the complexity of the experimental environment and increasing control by the experimenter. One of the actions is typically rewarded according to some reinforcement schedule, which may distribute food or aversive stimuli as reward or punishment, respectively, inducing instrumental learning (Yin & Knowlton, 2006). Additionally, the rate with which animals choose a specific action is recorded and provides a behavioral measure, see e.g. (Adams, 1982).

When put into the Skinner box, the animal first learns about its environment and uses this knowledge instrumentally and in a goal-directed manner. With prolonged training, the animal's response rate increases as behavior becomes more and more habitual and automatic (Yin &

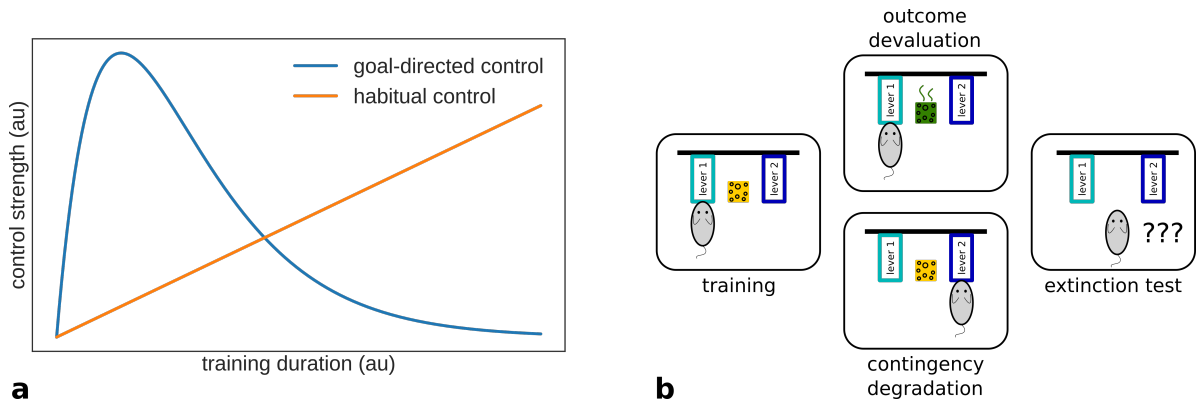

Figure 1.1: Acquisition of and probing for habitual behavior in animal experiments
a Qualitative control strengths of goal-directed (blue) and habitual (orange) control on behavior as a function of training duration, adapted from Figure 1B in (B. W. Balleine, 2019). The axes are in arbitrary units (au). Upon encountering a new environment, an agent typically learns about its structure and which behavior is rewarding. After having learned the action-outcome contingencies, the agent uses those to navigate to a goal in a goal-directed manner. As a result, the contributions of goal-directed control (blue) are thought to increase rapidly for short training durations (in typical experiments ≈10 trials). Once the agent has sufficiently learned appropriate behavior for its environment, goal-directed control strength peaks and continues to decline, as the agent can now rely on past rewards in order to control behavior, and does not need to engage in costly and slow explicit forward planning. Habitual control (orange) is thought to be low in the beginning and increases slowly with time, while the agent accumulates a tendency to repeat previously successful behavior. For moderate training durations (e.g. 50-100 trials), both control modes influence behavior. For extended training periods (e.g. 500 trials), the agent becomes over-trained, and typically almost solely relies on habitual control for action selection. Trial numbers were taken from (Colwill & Rescorla, 1988; Adams, 1982).
b A typical experimental setup to test for the strength of habitual control, adapted from Figure 1A in (B. W. Balleine, 2019). The agent, here a mouse, is put into a Skinner box (black box) with, for example, two levers. In the training phase (left box), usually one specific action, here pressing the left lever, is rewarded, while other behavior is not. After training, an experimental manipulation is applied (middle boxes), so that either the action-outcome contingencies change (contingency degradation, bottom box), or reward is devalued (outcome devaluation, top box). Afterwards, the continuation of previously learned behavior is tested in extinction (right box). If the agent is able to adapt its behavior to the changed environment, this is interpreted as evidence that the agent applies goal-directed control. If the agent is not able to adapt its behavior and continues to use previously learned behavior, it is interpreted as the agent applying mainly habitual control.

Knowlton, 2006; B. W. Balleine, 2019; Adams, 1982). In order to test if behavior has become habitual, the environment is then altered in a way that makes the previously reinforced action undesirable (Yin & Knowlton, 2006; B. W. Balleine, 2019). This is typically done using (i) contingency degradation or (ii) outcome devaluation (Figure 1.1**b**). Under contingency degradation, the action-outcome contingencies of the environment are changed by the experimenter, so that either abstaining from the action or a different action is rewarded. Under outcome devaluation, the reward magnitude is altered by the experimenter, by either satiating the animals so that food becomes less desirable, or even by causing an adverse reaction by, for example, taste aversion as a consequence of inducing a gastric illness. The response rate after this change is recorded to determine if and how long the animal continues to choose the previous action, which provides a measure of habit strength, for a review of these procedures see (Yin & Knowlton, 2006). Typically, the response rate and the time in which the old behavior is repeated increases with the duration of the training phase, while the reaction time decreases (Adams, 1982; Dickinson et al., 1983; Seger & Spiering, 2011). The habit strength is furthermore dependent on the specific action-outcome contingencies used and the reinforcement schedule which is chosen to deliver the rewards (Yin & Knowlton, 2006).

These results from the animal habit learning literature are often interpreted such that goal-directed behavior rests on action-outcome contingencies, while habits are cued stimulus-response associations (Smith & Graybiel, 2016; Yin & Knowlton, 2006). In this view, the cue can be implicit, such as the lever in the Skinner box, or more explicit, e.g. a light or tone is played to signal reward availability. Additionally, and as a generalization of the cued stimulus-response habits, some studies showed that habits are profoundly influenced by experimental context, e.g. (Bouton & Bolles, 1979; Thrailkill & Bouton, 2015) and can be quickly recalled when encountering a previously experienced context. Consequently, it has been argued that contexts may play an important role in instrumental learning (Bouton, 2019). In a similar vein, habit learning studies of humans in every-day behavior showed that habits are learned in a context-dependent manner (Lally, Van Jaarsveld, Potts, & Wardle, 2010; Wood & Rünger, 2016). Here, the context is often more implicitly cued, and equates to specific everyday-life situations such as "after breakfast". It has been found that habits are more easily learned when the behavior is repeated in the same context, e.g. always after breakfast (Lally et al., 2010; Danner, Aarts, & de Vries, 2008; D. T. Neal, Wood, Labrecque, & Lally, 2012), while breaking of habits is facilitated after a context switch, e.g. after a move to a new city (Verplanken & Roy, 2016).

To summarize, in experiments behavior is classified as habitual when it is insensitive to outcome devaluation and/or contingency degradation. Habit strength increases with training duration, and depends critically on action-outcome contingencies and reinforcement schedules used during training. Habitual behavior may be cued, and habit learning trajectories are context dependent, so that known behavior can be recalled quickly in a known environment. There have been several attempts to translate this operational definition into paradigms which induce habitual behavior in humans which will be discussed in the following.

Human experiments

One class of paradigms tried to directly translate outcome devaluation tests from animal experiments, see e.g. (Valentin, Dickinson, & O'Doherty, 2007; E. Tricomi, Balleine, & O'Doherty, 2009; Watson, Wiers, Hommel, & De Wit, 2014). In such tasks, participants are trained to

perform an action to obtain a reward, e.g. sweets, and are subsequently offered rewards until they are satiated. Afterwards, participants continue the task and it is measured whether they continue to perform the action despite the reward having lost its appetitive value. Using this kind of outcome devaluation paradigm, some studies were able to find inter-individual differences in sensitivity to outcome devaluation for different mental disorders compared to neurotypical participants, e.g. (Gillan et al., 2011; Everitt & Robbins, 2016). Nonetheless, de Wit et al. (2018) found that different variants of such outcome devaluation tasks failed to produce the effect that habit strength increases with increased training duration. These findings call into question whether such a simple single trial paradigm indeed induces habitual behavior in humans. Consequently, it may not be possible to map animal habit learning experiments to humans one to one, because humans may be too adaptive, or less easily trained (Watson & de Wit, 2018).

Another class of paradigms made use of more complex sequential decision making tasks, commonly using a sequence of two decisions to reach a goal (Daw, Gershman, Seymour, Dayan, & Dolan, 2011). This two-step task was set up such that two actions can be chosen and probabilistically lead to one of two states, where one transition is more likely than the other. Then, a second action could be chosen as one of two options, where a reward is given to the participant according to a probability which is evolving over time. The authors hypothesized that under goal-directed control, participants maintain a mental map of the task which they update based on the states visited and actions chosen. Contrarily, under habitual control, the authors proposed that participants may only update their representation of how good an action was, independent of the state they subsequently visited. If a rare transition occurred, and a state was visited which was unlikely given the action chosen, goal-directed and habitual control signals would be in contrast. This kind of task indeed induces stable and replicable effects, e.g. (Deserno et al., 2015; Otto, Raio, Chiang, Phelps, & Daw, 2013; Eppinger, Walter, Heekeren, & Li, 2013; Smittenaar, FitzGerald, Romei, Wright, & Dolan, 2013; Otto, Skatova, Madlon-Kay, & Daw, 2014), and the authors found evidence of neural correlates for both hypothesized control modes in fMRI experiments.

To investigate whether these hypothesized control modes can be mapped to the operational definition of habitual and goal-directed behavior, two studies combined instrumental learning in the two step task with a subsequent outcome devaluation (Friedel et al., 2014; Gillan, Otto, Phelps, & Daw, 2015). Interestingly, both studies found that the degree to which participants use the hypothesized habitual control mode in the two-step task did not correlate with the insensitivity participants showed under outcome devaluation. Both studies have limitations, namely low number of participants or being run on Amazon's Mechanical Turk, so they only present preliminary evidence that the constructs measured with the two-step task and outcome devaluation may not equate. Consequently, it is currently unclear whether the two-step task indeed measures habitual behavior as defined in the operational sense above, or another heuristic behavioral strategy. Further studies will need to investigate if the two-step task is a suitable paradigm to measure habit learning, or if other sequential decision making tasks better induce and measure habitual behavior.

1.2 Neural correlates of habit learning

Ideally, a mechanistic definition, as the one proposed in this thesis, must not only be congruent with the operational definitions of habits and their properties found in behavioral experiments, but should also allow for an interpretation in terms of brain function. Having discussed

operational and behavioral definitions of habits above, in this Section provides an overview about what is currently known about the neurobiological underpinnings of habitual and goal-directed behavior in rodents and humans.

In order to illustrate neurophysiological findings I present a simplified parcellation of cortico-basal ganglia-thalamo-cortical loops which have been found to be involved in implementing habitual as well as goal-directed behavior, see Figure 1.2. In accordance with the literature (B. W. Balleine, 2019; B. W. Balleine & O'Doherty, 2010; Yin & Knowlton, 2006), these loops can be divided into three sub-loops which are implicated in different stages of instrumental learning. Evidence suggests a strong interaction between the loops, which is not shown in this simplified overview.

It has been found that the loop centered on the dorsomedial striatum, the caudate nucleus in humans, is necessary for goal-directed behavior in the the early phases of instrumental learning (B. W. Balleine, 2019) when the agent learns action-outcome associations (middle loop in Figure 1.2). The prelimbic cortex in rodents, has been shown to recruit and coordinate the necessary components of goal-directed behavior: Working memory, synaptic plasticity (Tsutsui, Oyama, Nakamura, & Iijima, 2016); and sensory, affective, and motor areas (B. W. Balleine, 2019). Evidence suggests that it enables the dorsomedial striatum to hold learned action-outcome contingencies through its projections (Yin, Ostlund, Knowlton, & Balleine, 2005; E. M. Tricomi, Delgado, & Fiez, 2004). The medial prefrontal cortex has been suggested as a human homologue to the prelimbic cortex, where fMRI studies showed that its BOLD signal correlates with the expected reward of actions (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; B. W. Balleine & O'Doherty, 2010) and aides action value comparisons (Morris, Dezfouli, Griffiths, & Balleine, 2014). The dorsomedial striatum has been found to integrate inputs from prefrontal regions and the thalamus (Reep, Cheatwood, & Corwin, 2003; B. W. Balleine & O'Doherty, 2010) to encode action-outcome contingencies in a context- or state-dependent manner (Fino, Vandecasteele, Perez, Saudou, & Venance, 2018). The dorsomedial striatum is known to project its representation of learned behavior to the substantia nigra pars reticulata, an output nucleus of the basal ganglia, which in turn projects to the mediodorsal thalamus, whose disinhibition facilitates the initiation of movement in the motor areas (Pollack, 2001; B. W. Balleine & O'Doherty, 2010).

For habitual action on the other hand, evidence suggests much of its neural implementation to be based on a loop centering on the dorsolateral striatum (Reep et al., 2003; Yin, Knowlton, & Balleine, 2004), corresponding to the putamen in humans (right loop in Figure 1.2). It receives input from sensorimotor cortices, which have been shown to encode cues relevant for behavior (McGeorge & Faull, 1989) and thalamic regions (Alloway, Smith, Mowery, & Watson, 2017) which are thought to encode learned skills and unconditioned behavior such as orienting or whisking (B. W. Balleine, 2019). The dorsolateral striatum projects to the globus pallidus pars interna, another output nucleus of the basal ganglia, which projects to the ventral anterior and ventrolateral parts of the thalamus, again aiding the initiation of movement (Pollack, 2001). Compuatational studies furthermore suggested that habits are formed by chunking single actions into automatic sequences (Dezfouli & Balleine, 2012, 2013). This is supported by observed task bracketing activity in the dorsolateral striatum, where neurons are active at the beginning and end of an action sequence, and the magnitude of the activity is correlated with behavioral automaticity (Smith & Graybiel, 2013, 2014).

Additionally, there is thought to be a third loop in strong interaction with the two loops discussed above (left loop in Figure 1.2). It is centered on the ventral striatum (nucleus accumbens) and supports the two dorsal loops with retrieval of stored memory, and by providing incentive and subjective valuation (Yin & Knowlton, 2006; B. W. Balleine, 2019). For example,

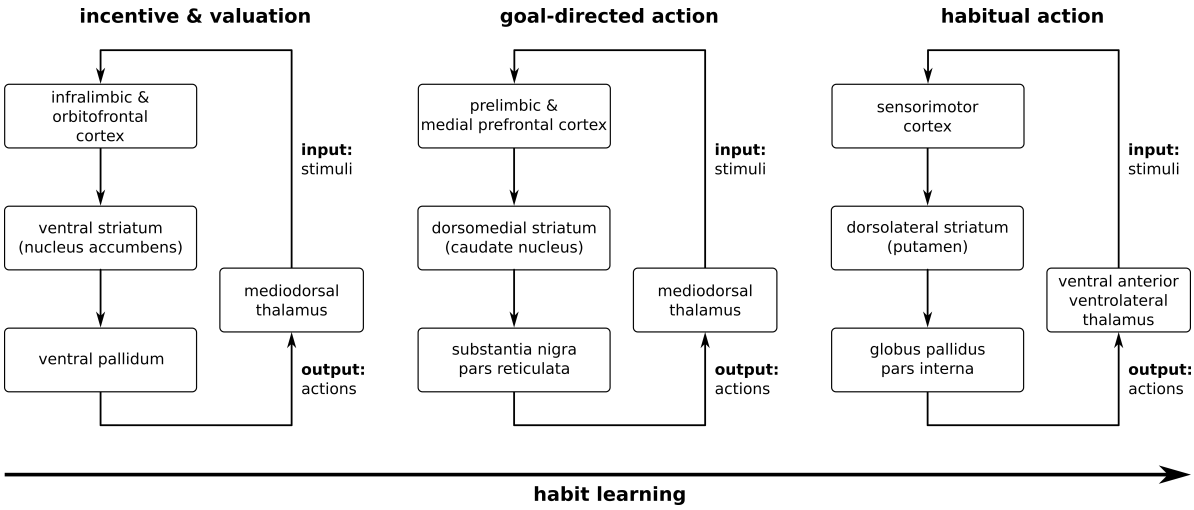

Figure 1.2: Cortico-basal ganglia-thalamo-cortical loops
This figure shows an approximate parcellation of the cortico-basal ganglia-thalamo-cortical loop into three sub-loops, adapted from Figure 3A in (B. W. Balleine, 2019), Figure 3 in (Yin & Knowlton, 2006). The left loop, centered on the ventral striatum, is has been found to be responsible for retrieval and valuation of action-outcome contingencies, and therewith is thought to mediate the incentives for different actions. The middle loop is centered on the dorsomedial striatum and has been shown to implement learning and on-line usage of action-outcome contingencies, which are necessary for goal-directed behavior. The right loop, centered on the dorsolateral striatum, has been found to process automatic, unconditioned, and habitual actions. All loops receive sensorimotor inputs through the thalamus, which allows for contextual retrieval, processing, and evaluation. As their output, they project back to the thalamus, where they aide motor preparation through disinhibition. Not shown are many cross connections between the three loops which also play a major role in action evaluation (see main text).

evidence suggests that the orbitofrontal cortex is involved with the retrieval, but not learning, of context-dependent action-outcome associations (Gremel & Costa, 2013; Parkes et al., 2018) which it supplies to the dorsomedial striatum for goal-directed evaluation. The relative predicted value of actions, which is needed in order to compare and decide between the outcomes associated with different actions, has been suggested to be evaluated in the ventral striatum, i.e. the nucleus accumbens in humans (Corbit & Balleine, 2015; Parkinson, Cardinal, & Everitt, 2000) and corresponds to the prospective evaluation needed for goal-directed action. Additionally, the nucleus accumbens has been found to also encode retrospective experienced value (B. W. Balleine, 2019) together with the amygdala (B. W. Balleine & Killcross, 2006). The ventral striatum projects to the ventral pallidum, which encodes subjective liking of a stimulus (Berridge & Kringelbach, 2015) which in turn projects to the mediodorsal thalamus.

As this description suggests, the habitual and goal-directed systems most likely are not competing as two opposing systems, but are rather intertwined and cooperate on the task of action evaluation (B. W. Balleine & O'Doherty, 2010; B. W. Balleine, 2019). During habit learning, a shift from the processing loop centering on the dorsomedial striatum to the loop centering on the dorsolateral striatum can be observed (Yin & Knowlton, 2006; B. W. Balleine, 2019). Additionally, human studies on addiction hint at a general shift from processing in the ventral striatum to processing in the dorsal striatum as addictive behavior is acquired (Everitt & Robbins, 2013), which is also understood as the behavior becoming more automatic and habitual.

1.3 Models of habit learning

Neither the psychological nor the animal literature have converged on a mechanistic definition of habit learning. To achieve a more mechanistic understanding and definition of habit learning, and to be able to understand brain function in more depth, behavioral and computational causes and requirements need to be formalized (Krakauer et al., 2017; Gomez-Marin et al., 2014; Cooper & Peebles, 2015; B. W. Balleine, 2019). One way to to formalize mechanistic and causal models, which has been proven as useful tool in disciplines like physics, is mathematical modeling. A mathematical model allows to specify causal relationships with high precision, and predictions of the mathematical model can be used to design new experiments to test the hypotheses planted into the model. Vice versa, new experimental results need to be fed into the model, so that it is able to describe a wider range of phenomena. In cognitive neuroscience and psychology, this approach has been termed "the combined computational-experimental approach".

In the past, influential computational habit learning models have been proposed, which allow to formalize concrete hypotheses for habit learning mechanisms, and to compare the hypotheses by fitting the models to experimental data. In this section I want to introduce some of these models and the hypotheses they are based on, some of which I will incorporate or extend in the model proposed in this thesis. Typically, a habit learning model must describe how action-outcome contingencies are learned and evaluated to produce a goal-directed control signal, and how habits are learned based on past experience. Importantly, a model has to furthermore propose a way in which relative control strengths of the two modes of operation are balanced when selecting an action. It should furthermore exhibit many if not most typical features of habitual behavior, such as increased habit strength with increased training duration.

Much evidence suggests that goal-directed behavior can be mathematically described using

a Markov decision process, while the picture for habitual action control is much less clear, for a review see (Wood & Rünger, 2016). First, I will introduce Markov decision processes in general and why they are thought to describe goal-directed behavior, and then go on to present different concrete implementations in computational cognitive models. Finally I will discuss different proposals to date on how to integrate habit learning into these models, and how they implement balancing of control contributions.

1.3.1 Goal-directed behavior as a Markov decision process

In a Markov decision process, it is assumed that an agent has to navigate through a state space, where different actions may lead to different states with an ascribed probability (Sutton & Barto, 1998a). Rewards may then be awarded when visiting a certain state, or by executing a certain action in a certain state. For the present work, I will focus on the former without loss of generality. From this description it is already clear, why a Markov decision process and goal-direct behavior, which rests on predictions based on learned action-outcome contingencies, share important features. If an agent knows its state, it can use a state transition matrix to predict future states and future rewards given any action it may choose, providing the ability to perform explicit forward planning. Furthermore, this section will specifically focus on sequences of actions, as found in sequential decision making, as it is the more general case on one hand, and on the other hand, experimental evidence suggests that habits are learned as sequences, see the discussion above and (Smith & Graybiel, 2013, 2014; Dezfouli & Balleine, 2012, 2013).

This Subsection aims to introduce the basic ideas and components of a Markov decision process, as well as the typical notation which will be used in Chapters 2 and 3. A Markov decision process rests on a state and reward space space, which encode possible states an agent may visit and rewards it may achieve. If the state and reward spaces are known, state transition and reward generation matrices can be specified, which form the mathematical basis of the time evolution of the process. If the spaces are unknown, the matrices can be learned by an agent from interacting with its environment (Sutton & Barto, 1998a). Figure 1.3**a** shows an example which will serve to introduce the spaces and matrices more concretely: Opening a door with a key. Here, behavior consists of several steps to reach the goal of opening the door, and this process may be automatized into a habit in everyday life. For this example, let us assume that opening the door is comprised of the following steps: You stand in front of the door, take out your key ring, look for the correct key, insert it into the lock, turn it, and open the door. The state space $\mathcal{S}$ is then defined by the set of five states

$$
\begin{aligned}
s^1 &= \text{being in front of the door} \\
s^2 &= \text{having the keys in your hand} \\
s^3 &= \text{having a specific key in your hand} \\
s^4 &= \text{having inserted the key into the lock} \\
s^5 &= \text{door is open}
\end{aligned}
$$

where four potential actions can be performed

$$
\begin{aligned}
a^1 &= \text{take keys out} \\
a^2 &= \text{look for key} \\
a^3 &= \text{insert key into lock} \\
a^4 &= \text{turn key.}
\end{aligned}
$$

Not all actions can be executed in all states. Lastly, the reward space $\mathcal{R}$ is defined by the set of two rewards

$$
\begin{aligned}
r^1 &= \text{no reward (door is closed)} \\
r^2 &= \text{reward (door is open)}
\end{aligned}
$$

see also Figure 1.3**a** for a graphical representation. State transitions can now be specified by defining which state can be reached from what state with a certain probability by taking a specific action, e.g. action a^1

$$
p\left(s'|s, a = a^1\right) = \begin{pmatrix} 0.1 & 0 & 0 & 0 & 0 \\ 0.9 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{1.1}
$$

Note that not all actions will reliably lead to a new state, for example, you may grab into your pocket for the keys, but if the keys are actually in a different pocket, you will not transition from the state of standing in front of the door to having the key ring in your hand (first column of the matrix). Furthermore, not all actions make sense to take in specific states, so they do not lead to a state transition (other columns).

Similarly, the reward generation matrix can be defined as

$$
p\left(r|s\right) = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{1.2}
$$

so that only the last state, when the door is open, is rewarded. This way, the problem of unlocking the door was defined as a sequential decision problem, and this process can also be unrolled in time, see Figure 1.3**b**. In this view, time-dependent states and rewards can be defined, and using the transition matrices can answer the question, in what state s_t is the agent at some arbitrary time step t, given it chose action a_{t-1}, and will it receive a reward r_t? In each time step, it can be in one of the 6 states defined above, so that $s_t \in \{s^1, \ldots, s^6\}$, it can receive one of the possible rewards ($r_t \in \{r^1, r^2\}$), and take one of the possible actions ($a_t \in \{a^1, \ldots, a^5\}$). In the rest of this work, I will focus on finite horizon Markov decision processes, which have a distinct start step at $t = 1$, and a last time step at $t = T$. This way, the agent can look T time steps into the future. The notation for the state transition and reward generation probabilities changes slightly, but still contain the same matrices defined above

$$
\begin{aligned}
p\left(s_{t+1}|s_t, a_t\right) &= p\left(s'|s, a\right) \\
p\left(r_t|s_t\right) &= p\left(r|s\right).
\end{aligned} \tag{1.3}
$$

Note that an episode of length T only entails $T - 1$ actions, see also Figure 1.3. Due to the sequential nature of the Markov decision process, it is common to not evaluate single actions

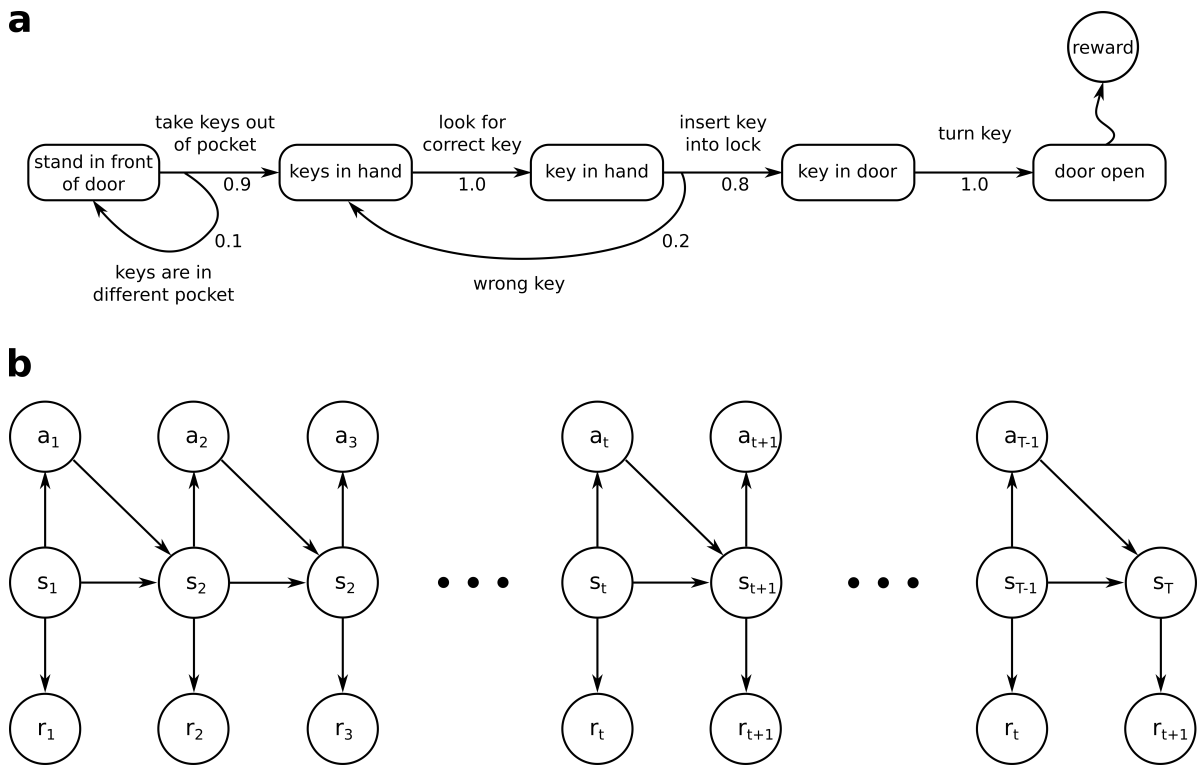

Figure 1.3: Markov decision process

Structural (**a**) and temporal (**b**) representation of a finite horizon Markov decision process for an exemplary planning process of opening a door with a key. **a** A graphical representation of the matrices defined in the main text showing states (boxes), actions (text), and state transitions (arrows) with their respective probabilities (numbers). **b** The Markov decision process unrolled in time. Circles indicate random variables in the decision process, arrows indicate statistical dependencies. The agent starts in the first time step $t = 1$ in state s_1. Depending on the state, a reward r_1 is generated. Note, that in this representation, a no-reward also is a possible outcome for r_t. Depending on the state the agent is in, it chooses an action a_1. In the next time step $t = 2$, the agent transitioned to a potentially new state s_2, in accordance with the previous action it took. Again, a reward r_2 is generated, and the agent chooses a new action a_2. This process repeats until the last time step $t = T$ is reached.

one by one, but to look at whole sequences of actions (Sutton & Barto, 1998a), which are typically called policies

$$\pi = (a_1, \ldots, a_{T-1}) \tag{1.4}$$

and span all $T-1$ actions within the planning horizon.

Having defined the Markov decision process and the matrices, the agent now has to ask the question, given my knowledge about the structure of the world, which action or policy should I choose to gain maximum reward? Computationally, this question can be answered in multiple ways, which rest on differing assumptions. One prominent way to solve such a Markov decision process is model-based reinforcement learning (Sutton & Barto, 1998a). Here, a value function containing the expected rewards for a policy is calculated. Usually, a decision rule is defined which transforms the values of policies into probabilities for actually following them. Another, emerging way to deal with Markov decision processes is planning as inference (Attias, 2003a; Botvinick & Toussaint, 2012a). Here, the matrices are treated as conditional probability distributions of a Bayesian generative model, where the (posterior) probability of choosing a specific policy is calculated using Bayesian inversion, see Section 1.4.2 for a more detailed description of this approach. A specific instance of planning as inference is the so-called active inference framework (K. Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty, & Pezzulo, 2016; K. Friston et al., 2015; Da Costa et al., 2020), where the Bayesian inversion is approximated using variational inference based on the minimization of the variational free energy. To summarize, model-based reinforcement learning, planning as inference, and active inference, all model goal-directed behavior based on a Markov decision process, despite their differences in how actions are being evaluated, which I will discuss below in more detail for Bayesian models of goal-directed behavior.

1.3.2 Approaches to modeling habit learning

For a computational description of habit learning, the evidence is much less clear what modeling approach should be used, and there have been many proposals which capture some properties of habit learning as discussed in Section 1.1. This section will introduce some of the most influential habit learning models, as well as models that rest on hypotheses which will be used in the proposed model, see Chapter 3. The discussion starts with one of the most published class of models, the model-free/model-based reinforcement learning models (Daw, Niv, & Dayan, 2005; Daw et al., 2011). It will go on to alternative modeling approaches based on reinforcement learning: A hierarchical approach to habit learning (Dezfouli & Balleine, 2012, 2013), as well as the value-free/value-based model (Miller, Shenhav, & Ludvig, 2019) where habits are learned based on repetition alone. This is followed by an introduction of three Bayesian approaches to modeling habit learning (Maisto, Friston, & Pezzulo, 2019; FitzGerald, Dolan, & Friston, 2014a; K. Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty, & Pezzulo, 2016), and lastly follow two context learning models (Gershman, Blei, & Niv, 2010; Redish, Jensen, Johnson, & Kurth-Nelson, 2007).

Model-free/model-based reinforcement learning approaches

To recapitulate, habits are formed slowly and are mainly viewed as retrospective, as they arise from repetition of behavior which has been successful in the past. Furthermore, they are fast to compute, but inflexible to changes in action-outcome contingencies. It has therefore long been assumed that habits can be viewed as a form of cached, i.e. experienced, action values.

An obvious choice to mathematically model cached action values is model-free reinforcement learning (Sutton & Barto, 1998a) which rests on a Pavlovian conditioning theory proposed by Rescorla and Wagner (1972). In its simplest form of temporal difference (TD) learning, a value of an action $Q(a)$ is being maintained and, depending on the reward received, updated after each trial. Interestingly, the action values in model-free reinforcement learning are an approximation to the value function of a Markov decision process under model-based reinforcement learning (Barto, Sutton, & Watkins, 1989). There are other, more advanced, model-free reinforcement learning algorithms, which learn state-action values $Q_t(s, a)$ such as the state–action–reward–state–action algorithm (SARSA) (Rummery & Niranjan, 1994) and Q-learning (Watkins, 1989).

Because of these properties, (Daw et al., 2005) proposed to view habitual behavior as model-free reinforcement learning, and goal-directed behavior as model-based reinforcement learning. Typically, in this kind of model, both systems are evaluated in parallel and an additional arbitration unit is used, which governs how goal-directed and habitual control contributions are balanced. Here, the authors suggested a Bayesian way to evaluate uncertainties on the two action evaluation systems and proposed that their contributions to action control may be weighted in accordance with the respective uncertainties of the two systems. To test this hypothesis in a human experiment, Daw et al. (2011) introduced the two-step sequential decision task, as described above, where they viewed model-free reinforcement learning as habitual control, and model-based reinforcement learning as goal-directed control. The control contributions are weighted according to some "model-based-ness" parameter, which the authors inferred from participant behavior. The authors studied the task in an fMRI experiment and could find model-free and model-based correlates in BOLD brain activation signals. This task and modeling approach has since found many applications, e.g. (Deserno et al., 2015; Otto et al., 2013; Eppinger et al., 2013; Smittenaar et al., 2013; Otto et al., 2014).

Another related approach for habit learning using model-free and model-based reinforcement learning is the "plan until habit" model by Keramati, Smittenaar, Dolan, and Dayan (2016). The authors assumed a different kind of balancing between model-free and model-based contributions: Given limited resources like time and working memory, agents may only employ a finite horizon Markov decision process with a planning depth T, for which they evaluate the decision tree. At the leave nodes of this cropped decision tree, the authors proposed agents should use model-free action values to estimate the remaining parts of the task which they do not explicitly evaluate. This proposal is particularly elegant as it does not need to rely on an additional arbitration unit. In a three-step sequential decision making task, the authors were able to show that human participants use a decreased planning horizon under time pressure.

Despite these theoretical and experimental advances into modeling habitual and goal-directed control, it is unclear whether model-free reinforcement learning sufficiently describes habit learning processes (Gillan et al., 2015; Friedel et al., 2014; Watson & de Wit, 2018). Using the same two-step task as (Daw et al., 2011), Friedel et al. (2014) and Gillan et al. (2015) subjected participants to an additional outcome devaluation test to explicitly probe for habitual control contributions. The model-free/model-based habit learning model was then fitted to participant data in the sequential decision task to determine model-based and model-free control contributions in participants. Both studies found that increased model-based contributions during the sequential decision task were correlated with an decreased failure to adapt to outcome devaluation, meaning, the more goal-directed someone was, the more they were able to adapt their behavior under outcome devaluation. However, they did not find a correlation between model-free control contributions and failure to adapt under

outcome devaluation, questioning the interpretation of model-free reinforcement learning as a model of habit learning. On a more technical note, Akam, Costa, and Dayan (2015) showed that small modifications to the task structure can lead to correlations of action values, which can bias model comparison so that model-free strategies could be classified as model-based in the data analysis. Additionally, model-free learning does not exhibit all properties of habit learning described above (Miller et al., 2019). Furthermore, having two systems evaluate the same situations in parallel but in a different manner would defy the purpose of quick and resource efficient habit learning and habitual action evaluation.

Alternative reinforcement learning modeling approaches

Dezfouli and Balleine (2012, 2013) for example showed that a hierarchical reinforcement learning model fits behavior from the same two-step task used by Daw et al. (2011) better than the original model. Here, the authors proposed to view habits not as model-free reinforcement learning, but as a chunking of actions into sequences in a hierarchical model-based reinforcement learning model, where an agent can then choose to either execute single actions or sequences. This choice is based on a cost function which evaluates the speed-accuracy trade-off between a fast sequence and more accurate but slower single action evaluations. Since sequences as well as single actions are integrated into the same model, this approach does not have to rely on an additional arbitration unit. The proposed model and the interpretation of habits as sequences fits well with the neurobiological findings of so-called task-bracketing activity, as found in e.g. (Smith & Graybiel, 2013).

Along a different line of reasoning, Miller et al. (2019) proposed to view habit learning as based on a simple and "value-free" repetition of previous behavior, analogous to skill learning. Conversely, they view goal-directed action evaluation as any evaluation based on reward value, meaning model-free as well as model-based reinforcement learning would constitute two different modes of this "value-based" learning. Here, the arbitration was again based on a separate arbitration unit, which evaluated the strengths of goal-directed and habitual control based on the variance of the predictions of the two systems. The authors were able to show in simulations that they can replicate many classical habit learning findings with their value-based/value-free habit learning model.

Bayesian approaches

Bayesian approaches to modeling habitual behavior differ from the reinforcement learning based approaches in three important ways: (i) the Markov decision process is solved using Bayesian inference (for details see Section 1.4), (ii) a posterior probability distribution over policies is calculated, which describes how likely it is that an agent will choose to follow specific policies, and (iii) due to Bayes' theorem, arbitration is usually automatically included in the inference process, without the need for modeling an additional arbitration unit.

As discussed above, behavioral action evaluation and selection can also be cast as a Bayesian inference problem. Given the success of some of the reinforcement learning based approaches, there have been attempts to integrate similar concepts into Bayesian frameworks, particularly active inference. Maisto et al. (2019) proposed a model where habits are understood as cached action values in the context of active inference. Here, instead of reevaluating the posterior probability of choosing an action at each repetition of the same task, an agent would cache previously calculated values of the posterior. Using simulations,

the authors showed that when behavior is tested in extinction, an agent cannot adapt its behavior until it infers that the environment changed sufficiently to warrant a reevaluation of the posterior, which produces habit-like effects of repeating previous behavior in the extinction phase.

Using a different interpretation of habit learning, FitzGerald et al. (2014a) investigated the model comparison aspect of the posterior over policies in a Bayesian framework, which takes over the role of balancing different control contributions. The authors show that a behavioral policy may be interpreted as a "model of behavior", and that any simpler policy based on a simpler model of behavior or of the environment, will be preferred in Bayesian model comparison when evaluating the posterior over actions. They stipulate that this may be a way in which habits emerge as a simpler model of behavior compared to the full evaluation of the Markov decision process. Using this approach, and combining it with the idea that habits can be understood as cached state-action values as in model-free reinforcement learning, e.g. SARSA, K. Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty, and Pezzulo (2016) proposed to view habits as a simpler state-action policy mapping states to actions which they added to an active inference model. The posterior over actions will then prefer the state-action policy due to its simplicity over the more complex evaluation of the Markov decision process, if the habit provides sufficiently desirable results. The authors showed in simulations that this approach yields habit-like effects under outcome devaluation.

Context models

Albeit being more focused on conditioning, extinction, and renewal than on habit learning per se, relevant models have been proposed which describe acquisition of behavior in the training phase, suppression in the extinction phase, as well as renewal when an agent is re-exposed to a known environment. Here, the training, extinction, and renewal phases of an experiment are interpreted as different contexts, where habit-like behavior in the extinction phase is due to context inference, see (Bouton, 2019) for a review. An influential model of this kind was proposed by Redish et al. (2007), who described the conditioning in the training phase using model-free temporal difference reinforcement learning. Additionally, the authors introduced context classification using a radial basis function network to identify the different phases of an experiment, which was able to guide state inference and learning. Along the same line of thought, Gershman et al. (2010) proposed a context model which used a particle filtering model instead of radial basis functions. Both proposals capture properties of habit learning by modeling the extinction and renewal phases of a conditioning experiment, but did not explicitly incorporate goal-directed behavioral control as e.g. in the form of a Markov decision process, therefore making them only partial habit learning models. Nonetheless, these are interesting proposals, as they incorporate key aspects of habit learning, which other models above are not able to describe, such as the context sensitivity and reinstatement of behavior.

1.4 Methods and modeling

Bayesian models are a compelling type of computational cognitive models, which will be used for the modeling approach proposed in this work. This Section will provide a short overview over the main ideas in Bayesian cognitive modeling, as well as planning as inference and active inference.

1.4.1 Bayesian cognitive models

The Bayesian brain hypothesis proposes a way to formalize the general information processing scheme in the brain using Bayesian probabilities and information theory (Knill & Pouget, 2004; Doya, Ishii, Pouget, & Rao, 2007; Clark, 2013). The basis for this information processing is the so-called generative model, a probabilistic model which encodes causality relationships and the time evolution of external variables and quantities of interest (Bishop, 2006a). Specifically, out of the many properties our environment may have, and states it could be in, only some are directly observable for an organism, like e.g. light reflections which can be detected by the retina. Other variables, like for example the identity of the object which caused the observed light reflections, are not directly observable by an agent, or are inherently unobservable, like future states and the time evolution of the environment. The generative model provides a way to formalize the rules according to which a hidden state, e.g. object identity, causes an observation, e.g. the corresponding light patterns falling on the retina, and how this hidden state might evolve with time. For exemplary purposes we can specify a simple generative model as

$$p\left(s,o\right)=p\left(o|s\right)p\left(s\right) \tag{1.5}$$

where the prior $p\left(s\right)$ defines an a priori knowledge about how often a specific hidden state s occurs, i.e. how often one would encounter a specific object, and the likelihood $p\left(o|s\right)$ encodes the rules according to which observations o, the light patterns, are generated from the hidden state. When an organism that encodes such a generative model now makes a certain observation, it needs to ask the inverse question in order to form beliefs about hidden states in the environment: What is the probability that a specific hidden state caused the observation that was just made? Formally, this equates to inverting the generative model according to Bayes' rule

$$p\left(s|o\right)=\frac{p\left(s,o\right)}{p\left(o\right)}=\frac{p\left(o|s\right)p\left(s\right)}{p\left(o\right)} \tag{1.6}$$

where the posterior $p\left(s|o\right)$ describes the inferred probability of a hidden state s given some observation o. Note that not only the likelihood, i.e. the rules, are used for this inference, but also the prior. This means that the inferred estimate will be biased towards states that were encountered more often in the past. While not yet generally accepted as a theory of brain function (for a critique see e.g. (Marcus & Davis, 2013; Kwisthout, Wareham, & van Rooij, 2011)), there is a growing body of evidence that human perception indeed works according to the principle of Bayesian inference (Clark, 2013). For example, Körding and Wolpert (2004) showed that participants learn a prior statistical distribution of a sensorimotor task and combine it with their sensory uncertainty in a Bayesian manner.

On the behavioral side, Attias (2003a) and Botvinick and Toussaint (2012a) proposed that planning can be cast as an inference process as well. Under this planning as inference idea, it is assumed that an agent not only uses Bayes' rule to infer the state of its surroundings, but also infers a probability distribution over actions or policies π, from which the current action that will be executed is sampled. Formally, this can be again described using Bayes' rule

$$p\left(\pi|r\right)=\frac{p\left(\pi\right)p\left(r|\pi\right)}{p\left(r\right)} \tag{1.7}$$

where the agent now infers the posterior $p\left(\pi|r\right)$ which policy π it should take given it wants to receive rewards r. This is calculated from the prior over policies $p\left(\pi\right)$, and the likelihood $p\left(r|\pi\right)$ of getting getting rewards when a policy is chosen. Thus, the likelihood encodes the

action-outcome contingencies and may be defined as a Markov decision process, as outlined in Section 1.3.1. This way, planning as inference offers a way to solve a Markov decision process in a Bayesian manner.

1.4.2 The free energy principle and active inference

One of the main challenges of the Bayesian brain hypothesis is that the Bayesian inversion can become computationally extremely costly, if not analytically intractable, for larger generative models (Brighton & Gigerenzer, 2008; Kwisthout & van Rooij, 2013), thus making it unlikely for the brain to analytically solve Bayesian inversion on the fly. As a remedy it has been proposed that the brain may use approximate Bayesian methods. One specific instance of such a proposal is the so-called free energy principle (K. Friston, 2009, 2010), which posits that the brain may use variational inference (Bishop, 2006a) to perform online inference efficiently. Here, an approximate posterior

$$\begin{aligned} q\left(s\right) &\approx p\left(s|o\right) \\ q\left(\pi\right) &\approx p\left(\pi|r\right) \end{aligned} \tag{1.8}$$

is used to approximate the true posteriors $p\left(s|o\right)$ and $p\left(\pi|r\right)$. The approximate posterior is typically assumed to be of a simpler form than the true posterior. Often, the so-called mean-field approximation is applied, which assumes statistical independence of hidden variables in the approximate posterior. Instead of using Bayes' theorem to calculate the true posterior, the variational free energy

$$F\left[q\right] = D_{KL}\left[q|p\right] \tag{1.9}$$

is used to calculate the approximate posterior, which is found at the minimum of this free energy (Bishop, 2006a). Hence, the free energy principle (K. Friston, 2009, 2010) offers a less costly way to calculate beliefs about hidden states and actions.

When calculating the beliefs over policies at the minimum of the free energy, the approximate posterior

$$q\left(\pi\right) \propto p\left(\pi\right) e^{-F(\pi)} \tag{1.10}$$

is calculated from a prior over policies $p\left(\pi\right)$ and the likelihood $e^{-F(\pi)}$, where $F\left(\pi\right)$ is the policy-specific free energy (K. Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty, & Pezzulo, 2016; K. Friston et al., 2015; Da Costa et al., 2020). This free energy encodes the goal-directed value of a policy, as it contains a prediction error from predicted to the desired outcomes, which is lower the more desirable the outcomes are. It is calculated from the underlying Markov decision process implemented into the generative model, which in turn contains action-outcome associations. In order to select an action, an agent samples a policy from the approximate posterior from which it executes the respective action in the sequence.

This way of solving a Markov decision process using approximate Bayesian methods as proposed in the free energy principle is also called active inference (K. Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty, & Pezzulo, 2016; K. Friston et al., 2015; Da Costa et al., 2020). Note that in most formulations of the active inference framework, the policy-specific free energy is called "expected free energy" and not only contains values from the Markov decision process, but also an epistemic value term, which gives additional value to actions which will yield new information, making them more likely to be chosen.

1.5 Open questions and hypotheses

As outlined in Section 1.3, there have been several attempts to propose computational and mechanistic accounts which were successful to varying degrees, where success may be measured by the model's ability to describe the key characteristics of habit learning

- habitual behavior under both contingency degradation and outcome devaluation,
- increased habit strength with increased training duration,
- why habits are more resource efficient than explicit forward planning,
- increased habit strength with decreased action-outcome contingency,
- context-sensitivity of habits and quick reinstatement of behavior in a familiar context,

and, importantly, by offering a sensible way to achieve a balancing of control, where the goal-directed evaluation does not have to run in parallel to the habitual evaluation. Most of the models did not attempt to emulate all properties, but often focused on one property in particular. For example, the model-free/model-based habit learning model set out to describe insensitivity to contingency degradation and outcome devaluation, but does not include an explicit description of contexts, and also lacks the resource efficiency for habitual behavior. Specifically, the model-based goal-directed evaluation has to be applied at every time step in order to find goal-directed control contributions, which is contrary to the very reason an agent should switch to habitual action control. Similar arguments can be made of all the of the models in Section 1.3, and none incorporates all these properties of habit learning, see Discussion for a detailed discussion.

Therefore, in this thesis, I want to propose a novel approach to mathematically model habit learning which exhibits all these properties in a unified framework. I want to build on the physiological and computational evidence that habits are context-dependent and expressed as sequences in a hierarchical model, and combine this proposal with that of value-free habit learning based on repetition of behavior. Therefore I hypothesize that

- habits are learned solely based on repetition of past behavior,
- habits can be viewed as action sequences embedded in a hierarchical model,
- habits are contextual and different habits are learned different contexts, along with the respective action-outcome contingencies,
- control contributions are based on the uncertainties of goal-directed and habitual evaluation.

To formalize these hypotheses, I will build a hierarchical Bayesian model where

- goal-directed evaluation is based on a Bayesian treatment of a Markov decision process,
- habits are learned as a prior over action sequences, which is updated each time a sequence has been chosen,
- control contributions are computed according to Bayes' rule in the posterior over actions from which an agent samples its action,

- context constitutes the upper level of the hierarchy,
- inference is based on approximate Bayesian computations.

Action selection, by sampling from the posterior, offers a simple and efficient way to balance respective control contributions. The posterior

$$p\left(\pi|r\right) \propto p\left(\pi\right) p\left(r|\pi\right) \tag{1.11}$$

describes the probability of choosing an action, given an agent wants to receive a reward. It is calculated from the prior $p\left(\pi\right)$, which corresponds to habits, and the likelihood $p\left(r|\pi\right)$ which evaluates the probability of attaining a reward based on the underlying Markov decision process. According to Bayes' rule, the two are weighted by a simple multiplication, which automatically achieves a balancing by the respective uncertainties of the two control contributions.

Additionally, since the prior has the capacity to exclude certain policies a priori, it also limits the amount of policies which will need to be evaluated in the goal-directed likelihood. This means, in this model, the two modes do not have to evaluate in parallel, but prior distribution over policies already constrains the evaluation the goal-directed system will undertake. In the case of a prior which only favors one specific policy, all other policies will be neglected and will not have to be evaluated. This fits well to the literature, where a habitual response can be executed quicker, and the process of a habit being interrupted by goal-directed values is slower.

In order to build such a model, I will use the active inference approach of using a Bayesian representation of the Markov decision process which I will solve using variational inference. I will start on a more technical note, and improve the policy inference process so that it works well with sequences. The most widely used mean-field approximation can yield erroneous inference when used on sequences of states, which I will show in Chapter 2 (Schwöbel, Kiebel, & Marković, 2018). To remedy this issue, the second order Bethe approximation can be applied, which allows for pairwise statistical dependencies in the approximate posterior, yielding improved inference on sequences and therewith improved goal-reaching performance of a simulated agent in a sequential decision task. In Chapter 3 I will go on to present the full hierarchical Bayesian habit learning model (Schwöbel, Markovic, Smolka, & Kiebel, 2019). I will show that an artificial agent based on that model exhibits the same behavioral patterns as animals in all classical habit learning settings. Lastly, in the Discussion I will summarize my findings, compare the proposed model to previous modeling approaches, and discuss implications of the proposed mechanistic definition of habits as well as limitations of the model in its current form.

2 Active inference, belief propagation, and the Bethe approximation

2.1 Abstract

When modelling goal-directed behavior in the presence of various sources of uncertainty, planning can be described as an inference process. A solution to the problem of planning as inference was previously proposed in the active inference framework in the form of an approximate inference scheme based on the variational free energy. However, this approximate scheme was based on the mean-field approximation, which assumes statistical independence of hidden variables and is known to show overconfidence and may converge to local minima of the free energy. To better capture spatio-temporal properties of an environment, we reformulated the approximate inference process using the so-called Bethe approximation. Importantly, the Bethe approximation allows for representation of pairwise statistical dependencies. Under these assumptions, the minimizer of the variational free energy corresponds to the belief propagation algorithm, commonly used in machine learning. To illustrate the differences between the mean-field approximation and the Bethe approximation, we have simulated agent behavior in a simple goal-reaching task with different types of uncertainties. Overall, the Bethe agent achieves higher success rates in reaching goal states. We relate the better performance of the Bethe agent to more accurate predictions about the consequences of its own actions. Consequently, active inference based on Bethe approximation extends the application range of active inference to more complex behavioral tasks.

2.2 Introduction

When trying to achieve goals, an acting agent typically lacks complete knowledge about its environment and is exposed to several sources of uncertainty in its environment. This makes the pursuit of goals a non-trivial task (Arthur, 1994; Simon, 1990).

Computational models for goal-directed behavior are typically based on the widely used computational framework of reinforcement learning (Sutton & Barto, 1998b) with a large

number of applications (Doll, Simon, & Daw, 2012; Rangel & Hare, 2010; Dayan & Niv, 2008; O'doherty et al., 2004; Montague, Hyman, & Cohen, 2004). However, a strong limitation of classical reinforcement learning models is that they do not take into account the influence of various sources of uncertainty on human behavior (Rushworth & Behrens, 2008; Doya, 2008; Behrens, Woolrich, Walton, & Rushworth, 2007; A. J. Yu & Dayan, 2005). Over the past years an increasing number of empirical findings have provided evidence that belief updating in humans closely follows that of a rational Bayesian agent (FitzGerald, Hämmerer, Friston, Li, & Dolan, 2017; Meyniel, Schlunegger, & Dehaene, 2015; Lake, Salakhutdinov, & Tenenbaum, 2015; Vossel et al., 2013; Payzan-LeNestour, Dunne, Bossaerts, & O'Doherty, 2013; Behrens, Hunt, Woolrich, & Rushworth, 2008; Daw et al., 2005). This suggests that humans actively use a representation of uncertainty when inferring the current and past states of the world, and when making decisions (K. Friston & Kiebel, 2009; T. S. Lee & Mumford, 2003a; Knill & Pouget, 2004; Dayan, Hinton, Neal, & Zemel, 1995).

In complex everyday environments decision making is affected by various sources of uncertainty hence in such settings it is useful to treat planning and action selection as an inference problem (Pearl, 1988; Attias, 2003b; Botvinick & Toussaint, 2012b; K. Friston et al., 2013). Under the *p*lanning as inference formulation, it is assumed that agents form beliefs (in a Bayes optimal manner) over possible future behaviors to decide upon the sequence of actions that allows them to reach their goals.

When modelling human decision making one typically postulates that the human brain uses an approximate inference scheme to update beliefs and generate plans (Mathys, Daunizeau, Friston, & Stephan, 2011; Nassar, Wilson, Heasly, & Gold, 2010; Daunizeau et al., 2010; Yuille & Kersten, 2006; Baker, Saxe, & Tenenbaum, 2006). Such an approximation is required to achieve computationally tractable and fast adjustments to behaviour in a dynamic environments (Nassar et al., 2010).

One approximate inference approach that is generically used in a wide range of applications is variational inference (Blei, Kucukelbir, & McAuliffe, 2017; Wainwright & Jordan, 2008; Beal, 2003; Bishop, 2006b). Variational inference also forms the formal basis of the *f*ree energy principle (K. Friston, 2010), which states that both action and perception underlie the minimization of the variational free energy of the past, current and the expected future sensations. As the variational free energy defines an upper bound on surprise (Bishop, 2006b; K. Friston, 2010), minimizing the free energy minimizes an agent's surprise about its sensations. In turn minimizing surprise improves an agent's representation of the environment and drives an agent to visit states from which the future is more predictable. This formulation was subsequently extended to model goal-directed behavior under uncertainty and is referred to as *a*ctive inference (K. Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty, & Pezzulo, 2016). In recent studies, active inference was successfully applied in the analysis of behavioral (K. Friston et al., 2014; Schwartenbeck et al., 2015) and neuroimaging data (Schwartenbeck, FitzGerald, & Dolan, 2016; Schwartenbeck, FitzGerald, Mathys, Dolan, & Friston, 2014).

Here we will revisit the variational treatment of planning as inference—motivated by core concepts of active inference—and provide step by step derivations of an active inference agent starting from basic definitions of planning as inference (Attias, 2003b; Botvinick & Toussaint, 2012b). Importantly, we will base the derivations on the so-called Bethe approximation (H. Bethe, 1931; H. A. Bethe, 1935), which will allow us to establish a formal link between the free energy principle and the set of update equations known as *b*elief propagation (K. J. Friston, Parr, & de Vries, 2017; Yedidia, Freeman, & Weiss, 2005; Pearl, 1988).

The standard approach for deriving an active inference agent is to base approximate inference on the so-called mean-field approximation (K. Friston, FitzGerald, Rigoli, Schwartenbeck,

O'Doherty, & Pezzulo, 2016). The key difference between the Bethe and the mean-field approximation lies in the way how approximate beliefs about trajectories are encoded. Technically, the mean-field approximation assumes that posterior beliefs about a sequence of states are approximated by a distribution in which beliefs over states are independent between time points. Crucially, this ignores the statistical dependencies inherent in state transitions, meaning that the approximate posterior estimates might converge to local optima of the free energy and exhibit over-confident belief representations throughout the decision making process (Weiss, 2001; Murphy, 2012).

For example, if I know that being in the state 1 will always result in a transition to state 2, then the surprise on moving from state 1 to state 3 can only be evaluated if I have a joint distribution over both states. This is precluded in the mean-field approximation but is retained in the Bethe approximation. This follows because the approximate posterior beliefs about any particular state are conditioned upon the previous state. Often, these pairwise statistical dependencies under the Bethe approximation even correspond to the true spatio-temporal dependencies of hidden states in a dynamic environment, so that the approximate posterior provides a tighter bound on the surprise, and hence exhibits less deviation from the exact posterior (Weiss, 2001). In principle, this means that any approximate Bayesian inference about trajectories in the past - or in the future - should be more accurate under a Bethe approximation, leading to more optimal behavior. For this reason the belief propagation algorithm is often applied in the machine learning field to sequential inference problems (Bishop, 2006b; Yedidia et al., 2005; S.-Z. Yu & Kobayashi, 2003; Fan, 2001; Rabiner, 1989; Gelb, 1974; Kalman, 1960).

In what follows we will provide a detailed, and rather didactic, technical overview of the basic elements needed to define planning as an inference problem, and relate its exact Bayesian solution to an approximate solution obtained using the variational approximation either under the mean-field or the Bethe approximation. To illustrate the approximation-dependent differences in goal-directed behavior in presence of uncertainty, we will introduce mean-field and Bethe based agents to a simple navigation task in a noisy grid world. Finally, using this proof-of-principle task we will show that an agent based on the Bethe approximation exhibits enhanced performance as compared to a mean-field based agent.

2.3 Methods

2.3.1 Generative process

In this paper we consider a sequential decision-making task, in which an agent executes a finite number of actions (choices) in order to reach a goal in a specific environment. Each choice is associated with a discrete time step $t \in [1, T]$, where T denotes the total number of time steps. Here we will model this decision process as a partially observable Markov decision process (Drake, 1962; Martin, 1967; Astrom, 1965; Monahan, 1982).

The task is defined as a 5-tuple $(H, A, \Theta, O, \Omega)$ (see Figure 2.1) where:

- H denotes a finite sized set of hidden states.
- A denotes a finite sized set of actions.
- Θ denotes action dependent conditional transition probabilities between states.
- O denotes a set of observations.

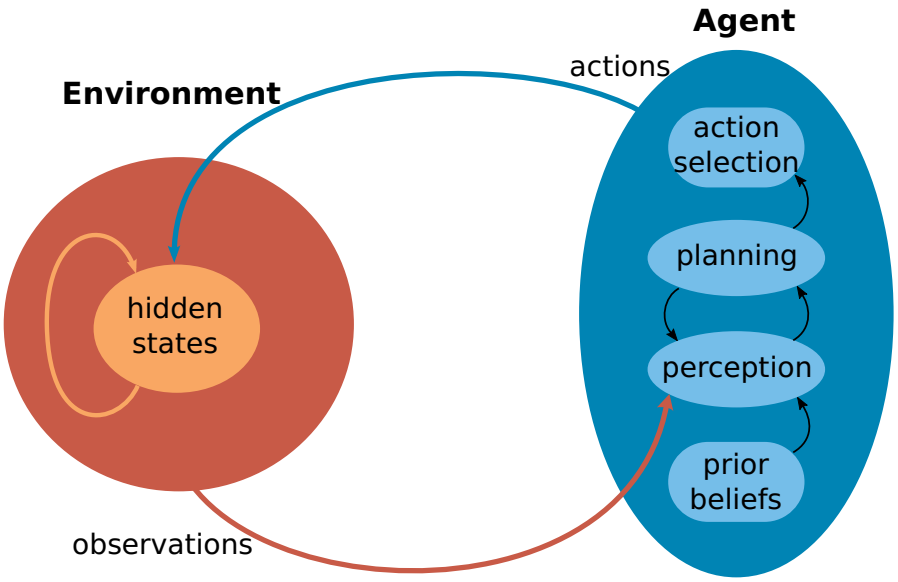

Figure 2.1: The environment and the active inference agent
The time evolution of the environment is defined via a generative process, which is conditionally dependent on the agent's actions. The agent can only indirectly access the hidden state of the environment via observations. The observations modulate the agent's beliefs about hidden states (perception), which in turn influence planning. Finally actions are selected to minimize surprise, that is, to fulfill the agent's prior beliefs about future observations, which encode the agent's goals. Importantly, the selected actions modulate the state transition process, hence influence the future state of the environment.

- Ω denotes state dependent conditional observation probability.

Each time step t of the generative process consists of the following components: Depending on the current state $\mathbf{h}_t \in H$, an observation $\mathbf{o}_t \in O$ is sampled from the generative probability $\Omega\left(\mathbf{o}_t|\mathbf{h}_t\right)$. Given an agents' choice of action $a_t \in A$ the environment will transit to a new state $\mathbf{h}_{t+1}$ sampled from the transition probability $\Theta\left(\mathbf{h}_{t+1}|\mathbf{h}_t, a_t\right)$. This process is repeated until the final time step T is reached.

2.3.2 Generative model

To efficiently solve the task, the agent needs an accurate representation of the generative process: The so-called *generative model* is a formal description of an agent's model of the hidden states of the environment and the rules that define their evolution. We will formally define the generative model as a joint probability distribution over observations $\mathbf{o}_t$, hidden states $\mathbf{h}_t$, and behavioral policies π, which define a sequence of control states u_t. Note that the control states denote a subjective abstraction of an action, e.g. a neuronal command to execute a specific action in the environment. For simplicity we will assume a one to one mapping between a selected control state u_t and executed action a_t in each time step t.

In line with previous definitions of a generative model used in behavioral models based on active inference (K. Friston et al., 2015; K. Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty, & Pezzulo, 2016), here we consider a special case in which each policy deterministically

Expression	Specification	Explanation
$\mathbf{h}_{1:T}$	$(\mathbf{h}_1, \ldots, \mathbf{h}_T)$	hidden states
$\mathbf{h}_t$	$\{h_1, \ldots, h_{n_h}\}$	current hidden state
$\underline{\mathbf{h}}$	$(\mathbf{h}_1, \ldots, \mathbf{h}_t)$	past (visited) hidden states, include current hidden state $\mathbf{h}_t$
$\tilde{\mathbf{h}}$	$(\mathbf{h}_{t+1}, \ldots, \mathbf{h}_T)$	future hidden states
$\mathbf{o}_{1:T}$	$(\mathbf{o}_1, \ldots, \mathbf{o}_T)$	observations
$\mathbf{o}_t$	$\{o_1, \ldots, o_{n_o}\}$	current observation
$\underline{\mathbf{o}}$	$(\mathbf{o}_1, \ldots, \mathbf{o}_t)$	past (fixed) observations, include current observation $\mathbf{o}_t$
$\tilde{\mathbf{o}}$	$(\mathbf{o}_{t+1}, \ldots, \mathbf{o}_T)$	future observations (unknown)
$u_{1:T-1}$	$(u_1, \ldots, u_{T-1})$	control states
u_t	$\{u_1, \ldots, u_{n_u}\}$	current control state
π	$u_{1:T-1}$	policy, a sequence of control states
$p(\mathbf{o}_{1:T}, \mathbf{h}_{1:T}, \pi)$		generative model, the agent's model of the rules of the environment
$\bar{p}(\tilde{\mathbf{o}})$		prior beliefs over future outcomes. These encode the agent's preference, or the utility of certain observations.
$f(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi \vert \underline{\mathbf{o}})$		true posterior, to be maximized
$q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi)$	$q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T} \vert \pi)\, q(\pi)$	approximate posterior
$q(\tilde{\mathbf{o}}, \mathbf{h}_{1:T} \vert \pi)$		agent's estimate of states and observations
$q(\pi)$	$\frac{1}{Z} p(\pi)\, \mathrm{e}^{-V_\pi - G_\pi}$	probability of following policy π
$F[q]$	$V[q] + G[q]$	Full variational free energy. Minimized by approximate posterior.
$V[q]$		observed free energy
V_π		conditional observed free energy under policy π
$G[q]$		predicted free energy
G_π		conditional predicted free energy under policy π

Table 2.1: Overview of the notation used in this article.

defines one possible sequence of control states. Conditioned on a behavioral policy $\pi = (u_1, \ldots, u_{T-1})$, we can express the full generative model over observations and hidden states as

$$p\left(\mathbf{o}_{1:T}, \mathbf{h}_{1:T}|\pi\right) = p\left(\tilde{\mathbf{o}}, \tilde{\mathbf{h}}|\mathbf{h}_t, \pi\right) p\left(\underline{\mathbf{o}}, \underline{\mathbf{h}}|\pi\right), \tag{2.1}$$

where the first factor on the right hand side

$$p\left(\tilde{\mathbf{o}}, \tilde{\mathbf{h}}|\mathbf{h}_t, \pi\right) = \prod_{\tau=t+1}^{T} p\left(\mathbf{o}_\tau|\mathbf{h}_\tau\right) p\left(\mathbf{h}_{\tau+1}|\mathbf{h}_\tau, \pi\right),$$

denotes the joint probability over future outcomes and hidden states, conditioned on a behavioral policy π. The second factor

$$p\left(\underline{\mathbf{o}}, \underline{\mathbf{h}}|\pi\right) = p\left(\mathbf{h}_1\right) \prod_{k=2}^{t} p\left(\mathbf{o}_k|\mathbf{h}_k\right) p\left(\mathbf{h}_k|\mathbf{h}_{k-1}, \pi\right),$$

denotes the joint probability over observed outcomes, and past hidden states. In practice we will derive the relations that define agent behavior (see Figure 2.1) by inverting the generative model. In what follows we describe in more detail the components of the full generative model. For the visualization of statistical dependencies between the random variables see Figure 2.2.

The agent's model of how the hidden state of the environment changes given a selected policy is formally expressed as

$$p\left(\mathbf{h}_{1:T}|\pi\right) = p\left(\mathbf{h}_1\right) \prod_{t=2}^{T} p\left(\mathbf{h}_t|\mathbf{h}_{t-1}, \pi\right). \tag{2.2}$$

where $p\left(\mathbf{h}_1\right)$ denotes the prior beliefs about the initial state $\mathbf{h}_1$, and $p\left(\mathbf{h}_t|\mathbf{h}_{t-1}, \pi\right)$ denotes an agent's beliefs about possible transitions between states, conditioned on the policy π. This conditional dependency is illustrated by the right pointing arrows in Figure 2.2. Note that each behavioral policy π defines a specific control state at each time step t, that is, $\pi(t) = u_t$. Hence the notation above is equivalent to replacing all π terms with the corresponding control states u_t at time step t.

Similarly, the agent requires a model of the relations between observations (outcomes) and hidden states of the environment

$$p\left(\mathbf{o}_{1:T}|\mathbf{h}_{1:T}\right) = \prod_{t=1}^{T} p\left(\mathbf{o}_t|\mathbf{h}_t\right). \tag{2.3}$$

Here, $p\left(\mathbf{o}_t|\mathbf{h}_t\right)$ denotes the conditional probability of making observation $\mathbf{o}_t$ in state $\mathbf{h}_t$ (this dependency is depicted by the arrows pointing downwards in Figure 2.2).

Given that the space of all possible behavioral policies $\pi \in \{1, \ldots, N_\pi\}$ is constrained by some prior distribution $p\left(\pi\right)$, we can write the simplified generative model as

$$p\left(\mathbf{o}_{1:T}, \mathbf{h}_{1:T}, \pi\right) = p\left(\pi\right) \prod_{k=2}^{T} p\left(\mathbf{o}_k|\mathbf{h}_k\right) p\left(\mathbf{h}_k|\mathbf{h}_{k-1}, \pi\right) p\left(\mathbf{h}_1\right). \tag{2.4}$$

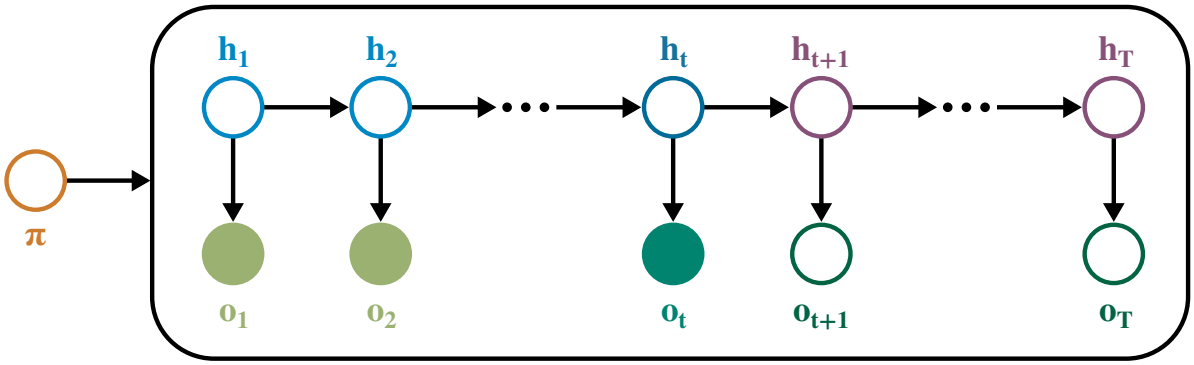

Figure 2.2: The full generative model as a Bayesian graph
Filled circles indicate observable (known/fixed) quantities, while the unfilled circles indicate hidden (unknown) variables. Arrows indicate the direction of conditional dependency between two variables. A policy π (brown circle) defines a specific sequence of control states which modulate state transitions and thereby the hidden states and observations. Light blue circles indicate past states $\underline{\mathbf{h}}$, dark blue the current state $\mathbf{h}_t$, and purple circles future states $\tilde{\mathbf{h}}$. Filled light and dark green circles depict past observations $\underline{\mathbf{o}}$, and the current observation $\mathbf{o}_t$ respectively. The green empty circles indicate future observations $\tilde{\mathbf{o}}$, which are also treated as hidden variables.

2.3.3 Planning as inference

The core concept of planning as inference is that besides the hidden states and future observations (see Figure 2.2), we treat the behavioral variables (control states, that is, policies) as hidden variables to be inferred (Attias, 2003b; Botvinick & Toussaint, 2012b). This approach has the advantage that these two different processes can be described within the same mathematical framework of Bayesian inference (Doya et al., 2007; Botvinick & Toussaint, 2012b). For this reason, the concept of describing planning as an inference process has found increasing interest within the cognitive neuroscience community (Botvinick & Toussaint, 2012b; Solway & Botvinick, 2012; K. J. Friston, Daunizeau, Kilner, & Kiebel, 2010).

Hence, as planning corresponds to computing the posterior joint probability over hidden states $\mathbf{h}_{1:T}$ and behavioral policies π, using Bayes rule we can write that

$$p\left(\mathbf{h}_{1:T}, \pi | \mathbf{o}_{1:T}\right) = \frac{p\left(\pi\right) \prod_{k=2}^{T} p\left(\mathbf{o}_k | \mathbf{h}_k\right) p\left(\mathbf{h}_k | \mathbf{h}_{k-1}, \pi\right) p\left(\mathbf{h}_1\right)}{p\left(\mathbf{o}_{1:T}\right)}. \tag{2.5}$$

The steps of the inference procedure can be illustrated as follows (see Figure 2.1): After making an observation, the agent updates its current beliefs about current and past (hidden) states $\underline{\mathbf{h}}$ (perception); from the inferred current state, the agents form beliefs about future states $\tilde{\mathbf{h}}$ and observations $\tilde{\mathbf{o}}$ for each policy π (planning).

Importantly the beliefs over policies (sequence of control states) are modulated by agent's preferences over unobserved future outcomes $\tilde{\mathbf{o}}$. We will represent these preferences as prior beliefs $\bar{p}\left(\tilde{\mathbf{o}}\right)$. Importantly, $\bar{p}\left(\tilde{\mathbf{o}}\right)$ defines the agent's goals, and thereby encodes the utility of various future outcomes (observations).

For example, if the goal is to reach a specific location (e.g. a position in a maze), the prior over future outcomes corresponds to assigning a high probability of observing an outcome specific of the goal location and low probability for other observations. Note that the prior beliefs over future outcomes are in general distinct from the marginal expectations over

future outcomes, that is, $\bar{p}(\tilde{\mathbf{o}}) \neq p(\tilde{\mathbf{o}})$. The difference is that these prior beliefs encode which observations the agent wants to make, while the marginal expectations represent where the agent will be at a future time step, given the time evolution of the environment.

As outlined above, in addition to the hidden states and policies we treat future outcomes $\tilde{\mathbf{o}}$ as hidden variables. Thus, we can express the complete joint posterior distribution as

$$f\left(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi | \underline{\mathbf{o}}\right) = p\left(\mathbf{h}_{1:T}, \tilde{u}, \pi | \tilde{\mathbf{o}}, \underline{\mathbf{o}}\right) \bar{p}(\tilde{\mathbf{o}}) \tag{2.6}$$

$$= \frac{p\left(\mathbf{o}_{1:T}, \mathbf{h}_{1:T}, \pi\right) \bar{p}(\tilde{\mathbf{o}})}{p\left(\mathbf{o}_{1:T}\right)} \tag{2.7}$$

Finally, we define the optimal policy, i.e. the optimal sequence of future actions as the mode of the posterior beliefs (Attias, 2003b)

$$\tilde{\mathbf{o}}^*, \mathbf{h}^*_{1:T}, \pi^* = \underset{\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi}{\arg\max}\, f\left(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi | \underline{\mathbf{o}}\right) \tag{2.8}$$

Once the agent computed which policy is optimal, it can choose an action accordingly (**a**ction selection).

In practice, as the generative model may represent an arbitrarily complex environment, inferring posterior beliefs over hidden variables is typically not analytically tractable (Bishop, 2006b). Therefore, to perform inference and select a policy, an agent would have to approximate the posterior beliefs (K. Friston & Kiebel, 2009; K. J. Friston et al., 2010).

2.3.4 Active inference

The active inference solution to the problem of planning as inference rests on variational inference. Typically, the variational free energy has been used under active inference for finding an approximate posterior distribution for the true posterior (Equation (2.6)) (K. J. Friston et al., 2010; K. Friston et al., 2013, 2015; K. Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty, & Pezzulo, 2016).

Variational free energy

Variational inference is a widely used approximate inference method (Blei et al., 2017; K. Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty, & Pezzulo, 2016; K. Friston et al., 2015, 2013; Wainwright & Jordan, 2008; Bishop, 2006b; Beal, 2003). For our particular problem of planning as inference, it will allow us to approximate the true posterior distribution $f\left(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi | \underline{\mathbf{o}}\right)$ with an approximate distribution $q\left(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi\right)$.

As a first step one has to define a set of potential candidate distributions for the approximate posterior, e.g. by constraining the posterior to a specific family of distributions. The best approximation to the true posterior is obtained as the distribution that minimizes the Kullback-Leibler (KL) divergence between the approximate and the true posterior, hence

$$q^* = \underset{q}{\arg\min}\, D_{KL}\left(q || f\right). \tag{2.9}$$

However, as the true posterior is not known a priori, the KL divergence cannot be minimized directly. However, if we substitute Equation (2.6) into Equation (2.9) we obtain the following expression

$$D_{KL}\left(q || f\right) = F[q] + \sum_{\tilde{\mathbf{o}}} q(\tilde{\mathbf{o}}) \ln p\left(\mathbf{o}_{1:T}\right), \tag{2.10}$$

where $F[q]$ denotes variational free energy defined as

$$\begin{aligned} F[q] &= \sum_{\tilde{\mathbf{o}},\mathbf{h}_{1:T},\pi} q\left(\tilde{\mathbf{o}},\mathbf{h}_{1:T},\pi\right) \ln \frac{q\left(\tilde{\mathbf{o}},\mathbf{h}_{1:T},\pi\right)}{p\left(\mathbf{o}_{1:T},\mathbf{h}_{1:T},\pi\right)\bar{p}\left(\tilde{\mathbf{o}}\right)} \\ &= -\sum_{\tilde{\mathbf{o}}} q\left(\tilde{\mathbf{o}}\right) \ln \bar{p}\left(\tilde{\mathbf{o}}\right) \\ &\quad - \sum_{\tilde{\mathbf{o}},\mathbf{h}_{1:T},\pi} q\left(\tilde{\mathbf{o}},\mathbf{h}_{1:T},\pi\right) \ln p\left(\mathbf{o}_{1:T},\mathbf{h}_{1:T},\pi\right) \\ &\quad + \sum_{\tilde{\mathbf{o}},\mathbf{h}_{1:T},\pi} q\left(\tilde{\mathbf{o}},\mathbf{h}_{1:T},\pi\right) \ln q\left(\tilde{\mathbf{o}},\mathbf{h}_{1:T},\pi\right). \end{aligned} \tag{2.11}$$

Given that the KL divergence is a positive quantity which goes to zero only for $q = f$, and that the variational free energy can be expressed as

$$F[q] = D_{KL}\left(q||f\right) - \sum_{\tilde{\mathbf{o}}} q\left(\tilde{\mathbf{o}}\right) \ln p\left(\mathbf{o}_{1:T}\right), \tag{2.12}$$

we get the following inequality

$$F[q] \geq -\ln p\left(\underline{\mathbf{o}}\right) - \sum_{\tilde{\mathbf{o}}} q\left(\tilde{\mathbf{o}}\right) \ln p\left(\tilde{\mathbf{o}}|\underline{\mathbf{o}}\right). \tag{2.13}$$

Hence, minimizing the variational free energy lowers the upper bound on the observed surprise $-\ln p\left(\underline{\mathbf{o}}\right)$, the future expected surprise $-\sum_{\tilde{\mathbf{o}}} q\left(\tilde{\mathbf{o}}\right) \ln p\left(\tilde{\mathbf{o}}|\underline{\mathbf{o}}\right)$, and minimizes the KL divergence between the true and approximate posterior. Thus, we can rewrite Equation (2.9) as

$$q^* = \arg\min_q F\left[q\right]. \tag{2.14}$$

Importantly, in the limiting case of $q = f$ the above inequality Equation (2.13) turns into an equality, that is,

$$F[q] = -\ln p\left(\underline{\mathbf{o}}\right) - \sum_{\tilde{\mathbf{o}}} \bar{p}\left(\tilde{\mathbf{o}}\right) \ln p\left(\tilde{\mathbf{o}}|\underline{\mathbf{o}}\right).$$

In accordance with Equation (2.11), the free energy $F[q]$ can be defined as a sum of two terms

$$F[q] = V[q] + G[q], \tag{2.15}$$

where we use $V[q]$ to denote the observed free energy

$$V[q] = \sum_{\underline{\mathbf{h}},\pi} q\left(\underline{\mathbf{h}},\pi\right) \ln \frac{q\left(\underline{\mathbf{h}},\pi\right)}{p\left(\underline{\mathbf{o}},\underline{\mathbf{h}},\pi\right)},$$

and $G[q]$ to denote the predicted free energy

$$G[q] = -\sum_{\tilde{\mathbf{o}}} q\left(\tilde{\mathbf{o}}\right) \ln \bar{p}\left(\tilde{\mathbf{o}}\right) + \sum_{\tilde{\mathbf{o}},\mathbf{h}_{1:T},\pi} q\left(\tilde{\mathbf{o}},\mathbf{h}_{1:T},\pi\right) \ln \frac{q\left(\tilde{\mathbf{o}},\tilde{\mathbf{h}}|\underline{\mathbf{h}},\pi\right)}{p\left(\tilde{\mathbf{o}},\tilde{\mathbf{h}}|\mathbf{h}_t,\pi\right)}.$$

In general we can express the approximate posterior q as a product of two factors: The marginal beliefs over policies $q\left(\pi\right)$ and the conditional beliefs over the remaining hidden variables $q\left(\tilde{\mathbf{o}},\mathbf{h}_{1:T}|\pi\right)$, that is,

$$q\left(\tilde{\mathbf{o}},\mathbf{h}_{1:T},\pi\right) = q\left(\tilde{\mathbf{o}},\mathbf{h}_{1:T}|\pi\right) q\left(\pi\right). \tag{2.16}$$

This allows us to find the minimizer of the variational free energy with respect to the marginal posterior over policies as

$$\frac{\delta F[q]}{\delta q(\pi)} \equiv 0,$$

which is obtained for

$$q(\pi) = \frac{p(\pi)\,\mathrm{e}^{-V_\pi - G_\pi}}{\sum_\rho p(\rho)\,\mathrm{e}^{-V_\rho - G_\rho}}, \tag{2.17}$$

where

$$V_\pi = \sum_{\underline{\mathbf{h}}} q\left(\underline{\mathbf{h}}|\pi\right) \ln \frac{q\left(\underline{\mathbf{h}}|\pi\right)}{p\left(\underline{\mathbf{o}}, \underline{\mathbf{h}}|\pi\right)}, \tag{2.18}$$

$$\begin{aligned} G_\pi = &- \sum_{\tilde{\mathbf{o}}} q\left(\tilde{\mathbf{o}}|\pi\right) \ln \bar{p}\left(\tilde{\mathbf{o}}\right) \\ &+ \sum_{\tilde{\mathbf{o}}, \mathbf{h}_{1:T}} q\left(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}|\pi\right) \ln \frac{q\left(\tilde{\mathbf{o}}, \tilde{\mathbf{h}}|\underline{\mathbf{h}}, \pi\right)}{p\left(\tilde{\mathbf{o}}, \tilde{\mathbf{h}}|\mathbf{h}_t, \pi\right)}, \end{aligned} \tag{2.19}$$

denote the conditional free energy of the past and of the future, respectively.

Note that in previous definitions of active inference (K. Friston et al., 2015; K. Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty, & Pezzulo, 2016; K. Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016), the approximate posterior over policies $q(\pi)$ was not defined as the minimizer of the full free energy. Instead, a prior over policies was defined as $\ln p(\pi) = G_\pi^{\text{expected}}$, with the so called *expected free energy* G_π^{expected}, from which the posterior over policies was derived. In the present formulation, the free energy driving agent behavior is not the expected free energy, but the conditional full free energy that allows us to express the approximate posterior using the sum of the conditional free energy from past observations plus the conditional free energy of the future. We call the conditional free energy of the past V_π *observed free energy*. It constrains the posterior beliefs over policies π to only those policies that could have generated the observed sequence given the agent's generative model. We refer to the conditional free energy of the future G_π as *predicted free energy*. The predicted free energy will be the main factor influencing policy selection. Here, the first term corresponds to pragmatic value or extrinsic value; namely the (negative), expected utility or log preferences over outcomes

$$E_q\left[\ln \bar{p}\left(\tilde{\mathbf{o}}\right)\right] = E_q\left[U\left(\tilde{\mathbf{o}}\right)\right].$$

The second term can be regarded as a consistency term. When it is minimized it ensures that the posterior beliefs about the future adhere to the generative model. The predicted free energy, as used in the present work, lacks the epistemic or ambiguity reducing component of the expected free energy. This component is usually associated with epistemic value or intrinsic value (also known as information gain or expected Bayesian surprise). It gives rise to altered agent behavior when compared to behavior chosen in accordance with the predicted free energy, which we will discuss later. For a more detailed insight into the differences of the two formulations and their relationship, we refer the reader to the appendix.

In what follows we will derive the update equations for the conditional posterior $q\left(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}|\pi\right)$ for two different approximations, the mean-field and the Bethe approximation.

Mean-field approximation

The mean-field approximation is a widely used approximation because of its simplicity, as it is based on the assumption of statistical independence of hidden variables (Bishop, 2006b). As in previous formulations of active inference (K. Friston et al., 2013, 2015; K. Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty, & Pezzulo, 2016), we assume statistical independence of the hidden states $\mathbf{h}_k$ and write the approximate posterior as

$$q\left(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}|\pi\right) = \prod_{\tau=t+1}^{T} q\left(\mathbf{o}_\tau|\mathbf{h}_\tau\right) \prod_{k=1}^{T} q\left(\mathbf{h}_k|\pi\right), \tag{2.20}$$

Inserting the ansatz Equation (2.20) into Equation (2.18) and Equation (2.19) yields the following relations for the conditional observed and predicted free energies

$$\begin{aligned}
V_\pi &= \sum_{r=1}^{t} V_\pi\left(r\right) && (2.21)\\
V_\pi\left(r\right) &= \sum_{\mathbf{h}_r, \mathbf{h}_{r-1}} q\left(\mathbf{h}_r|\pi\right) q\left(\mathbf{h}_{r-1}|\pi\right) \ln \frac{q\left(\mathbf{h}_r|\pi\right)}{p\left(\mathbf{o}_r|\mathbf{h}_r\right) p\left(\mathbf{h}_r|\mathbf{h}_{r-1}, \pi\right)} \\
G_\pi &= \sum_{\tau=t+1}^{T} G_\pi\left(\tau\right) && (2.22)\\
G_\pi\left(\tau\right) &= \sum_{\mathbf{o}_\tau, \mathbf{h}_\tau} q\left(\mathbf{o}_\tau, \mathbf{h}_\tau|\pi\right) \left[-\ln \bar{p}\left(\mathbf{o}_\tau\right) + \ln \frac{q\left(\mathbf{o}_\tau|\mathbf{h}_\tau\right)}{p\left(\mathbf{o}_\tau|\mathbf{h}_\tau\right)}\right] \\
&\quad + \sum_{\mathbf{h}_\tau, \mathbf{h}_{\tau-1}} q\left(\mathbf{h}_\tau|\pi\right) q\left(\mathbf{h}_{\tau-1}|\pi\right) \ln \frac{q\left(\mathbf{h}_\tau|\pi\right)}{p\left(\mathbf{h}_\tau|\mathbf{h}_{\tau-1}, \pi\right)}
\end{aligned}$$

The *u*pdate equations for the approximate conditional posterior are obtained as the minimizer of the conditional free energy

$$F_\pi = V_\pi + G_\pi$$

with respect to the factors of the approximate posterior $q\left(\mathbf{h}_k|\pi\right)$ and $q\left(\mathbf{o}_\tau|\mathbf{h}_\tau\right)$. It is important to note that only the first term of the predicted free energy, namely the cross entropy $-\sum_{\mathbf{o}_\tau, \mathbf{h}_\tau} q\left(\mathbf{o}_\tau|\pi\right) \ln \bar{p}\left(\mathbf{o}_\tau\right)$, will have a substantial influence on $q\left(\pi\right)$ and therewith goal-directed behavior. In other words, it is this term that constitutes the extrinsic or pragmatic value that maximizes the predicted log preferences. The remaining terms ensure that beliefs about future states conform to the known rules that govern the dynamics of hidden states, and known relations between the hidden states and sensory observations.

The resulting update equations are

$$\begin{aligned}
q\left(\mathbf{o}_\tau|\mathbf{h}_\tau\right) &= \frac{\bar{p}\left(\mathbf{o}_\tau\right) p\left(\mathbf{o}_\tau|\mathbf{h}_\tau\right)}{Z_\tau\left(\mathbf{h}_\tau\right)} \\
q\left(\mathbf{h}_k|\pi\right) &= \frac{m^k\left(\mathbf{h}_k\right) m_\pi^{k-1}\left(\mathbf{h}_k\right) m_\pi^{k+1}\left(\mathbf{h}_k\right)}{Z_k} && (2.23)\\
q\left(\pi\right) &= \frac{p\left(\pi\right) \mathrm{e}^{-G_\pi - V_\pi}}{\sum_\rho p\left(\rho\right) \mathrm{e}^{-G_\rho - V_\rho}}
\end{aligned}$$

where with m we denote various massages to yield comparability in notation with the following section. The messages are defined as

$$\begin{aligned} m^k\left(\mathbf{h}_k\right) &= \begin{cases} Z_\tau\left(\mathbf{h}_\tau\right), \text{ for } k > t \\ p\left(\mathbf{o}_k|\mathbf{h}_k\right), \text{ for } k \leq t \end{cases} \\ m_\pi^{k+1}\left(\mathbf{h}_k\right) &= \mathrm{e}^{\sum_{\mathbf{h}_{k+1}} q\left(\mathbf{h}_{k+1}|\pi\right) \ln p\left(\mathbf{h}_{k+1}|\mathbf{h}_k,\pi\right)} \\ m_\pi^{k-1}\left(\mathbf{h}_k\right) &= \mathrm{e}^{\sum_{\mathbf{h}_{k-1}} q\left(\mathbf{h}_{k-1}|\pi\right) \ln p\left(\mathbf{h}_k|\mathbf{h}_{k-1},\pi\right)} \end{aligned} \tag{2.24}$$

Note that the conditional posterior $q\left(\mathbf{h}_k|\pi\right)$ depends on the posterior beliefs at the neighboring time points $q\left(\mathbf{h}_{k+1}|\pi\right)$ and $q\left(\mathbf{h}_{k-1}|\pi\right)$. The optimal solution for the approximate posterior is obtained by iterating through Equation (2.23) and Equation (2.24) until convergence is achieved, as using Equation (2.23) directly leads to several practical problems. To ensure numerical stability and convergence of the update equations one typically resorts to the following gradient descent procedure (K. Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016)

$$\begin{aligned} x_\pi^{n+1,k} &= x_\pi^{n,k} + \epsilon\left(\rho_\pi^{n,k} - x_\pi^{n,k}\right) \\ q^{n+1}\left(\mathbf{h}_k|\pi\right) &= \frac{\mathrm{e}^{x_\pi^{n+1,k}}}{\sum_j \mathrm{e}^{x_\pi^{n+1,j}}} \\ \rho_\pi^{n,k} &= \ln m^k\left(\mathbf{h}_k\right) + \ln m_\pi^{n,k+1}\left(\mathbf{h}_k\right) + \ln m_\pi^{n,k-1}\left(\mathbf{h}_k\right) \end{aligned} \tag{2.25}$$

where we set the following initial conditions for each time step k.

$$x_\pi^{0,k} = \frac{1}{n_h}, \forall k \in \{1, \ldots, T\}, \text{ and } \forall \pi \in \{1, \ldots, N_\pi\}. \tag{2.26}$$

Bethe approximation

Under the mean-field approximation, statistical independence of hidden variables was assumed. This has the advantage of simplicity, as it makes it possible to analytically calculate the approximate posterior directly from the full free energy. When performing a sequential decision-making task, however, hidden states of the environment are most likely not independent of each other, but instead the current hidden state might depend on the previous hidden state. In other words, if the environment has a sequential structure, the mean-field approximation may not be able to capture this structure accurately. To address this issue of representing a sequential structure within the approximate posterior, the Bethe approximation (Pearl, 1988; Yedidia, Freeman, & Weiss, 2001b) can be used, as it allows for pairwise statistical dependencies between hidden variables in the approximate posterior. These dependencies map closely to the true statistical dependencies present in the generative model (see Figure 2.2).

For this reason, the Bethe approximation has found wide spread-usage in the machine learning community (Felzenszwalb & Huttenlocher, 2006; Coughlan & Ferreira, 2002; Sudderth, Mandel, Freeman, & Willsky, 2004; Hua, Yang, & Wu, 2005; Meltzer, Yanover, & Weiss, 2005). Using this more complex approximate posterior, the variational free energy becomes more complex to evaluate as well. In the past, it was shown that the estimation of the approximate posterior under the Bethe approximation corresponds to the *b*elief propagation update rules

(Pearl, 1988; Yedidia et al., 2001b). Belief propagation provides a framework to calculate the posterior beliefs using messages which are sent between nodes of the graph of the generative model. This solution using message passing provides the exact solution on a graph without loops, making the solution always converge to the global minimum of the variational free energy. For a detailed overview of belief propagation, the Bethe approximation and their relation to the variational free energy we point the reader to (Yedidia, Freeman, & Weiss, 2003a).

Under the Bethe approximation we express the functional form of the approximate conditional posterior as

$$q\left(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}|\pi\right) = \prod_{\tau=t+1}^{T} \frac{q\left(\mathbf{o}_\tau, \mathbf{h}_\tau|\pi\right)}{q\left(\mathbf{h}_\tau|\pi\right)} \prod_{k=1}^{T} \frac{q\left(\mathbf{h}_k, \mathbf{h}_{k-1}|\pi\right)}{q\left(\mathbf{h}_{k-1}|\pi\right)}, \tag{2.27}$$

where $q\left(\mathbf{h}_1, \mathbf{h}_0|\pi\right) = q\left(\mathbf{h}_1|\pi\right)$, and $q\left(\mathbf{h}_0|\pi\right) = 1$. Inserting Equation (2.27) for the approximate posterior in (2.18) and (2.19) we get the following form for the conditional observed and predicted free energies

$$V_\pi = \sum_{r=1}^{t} V_\pi\left(r\right) \tag{2.28}$$

$$V_\pi\left(r\right) = \sum_{\mathbf{h}_r, \mathbf{h}_{r-1}} q\left(\mathbf{h}_r, \mathbf{h}_{r-1}|\pi\right) \ln \frac{q\left(\mathbf{h}_r|\mathbf{h}_{r-1}, \pi\right)}{p\left(\mathbf{o}_r|\mathbf{h}_r\right) p\left(\mathbf{h}_r|\mathbf{h}_{r-1}, \pi\right)}$$

$$G_\pi = \sum_{\tau=t+1}^{T} G_\pi\left(\tau\right) \tag{2.29}$$

$$\begin{aligned} G_\pi\left(\tau\right) = & -\sum_{\mathbf{o}_\tau} q\left(\mathbf{o}_\tau|\pi\right) \ln \bar{p}\left(\mathbf{o}_\tau\right) \\ & + \sum_{\mathbf{o}_\tau, \mathbf{h}_\tau} q\left(\mathbf{o}_\tau, \mathbf{h}_\tau|\pi\right) \ln \frac{q\left(\mathbf{o}_\tau|\mathbf{h}_\tau, \pi\right)}{p\left(\mathbf{o}_\tau|\mathbf{h}_\tau\right)} \\ & + \sum_{\mathbf{h}_\tau, \mathbf{h}_{\tau-1}} q\left(\mathbf{h}_\tau, \mathbf{h}_{\tau-1}|\pi\right) \ln \frac{q\left(\mathbf{h}_\tau|\mathbf{h}_{\tau-1}, \pi\right)}{p\left(\mathbf{h}_\tau|\mathbf{h}_{\tau-1}, \pi\right)} \end{aligned}$$

As under the mean-field approximation, here the main contributing term for goal-directed behavior is the cross entropy $-\sum_{\mathbf{o}_\tau} q\left(\mathbf{o}_\tau|\pi\right) \ln \bar{p}\left(\mathbf{o}_\tau\right)$ in the predicted free energy, while the other terms ensure optimal posterior beliefs for the hidden states $q\left(\mathbf{h}_k|\pi\right)$ and future observations $q\left(\mathbf{o}_\tau|\pi\right)$.

To find the minimizer of the conditional free energy $F_\pi = V_\pi + G_\pi$ under the Bethe approximation we have to take into account the following equality constraints

$$\begin{aligned} q\left(\mathbf{h}_k|\pi\right) &= \sum_{\mathbf{h}_{k+1}} q\left(\mathbf{h}_{k+1}, \mathbf{h}_k|\pi\right) \\ &= \sum_{\mathbf{h}_{k-1}} q\left(\mathbf{h}_k, \mathbf{h}_{k-1}|\pi\right) \\ &= \sum_{\mathbf{o}_k} q\left(\mathbf{o}_k, \mathbf{h}_k|\pi\right), \\ q\left(\mathbf{o}_k|\pi\right) &= \sum_{\mathbf{h}_k} q\left(\mathbf{o}_k, \mathbf{h}_k|\pi\right). \end{aligned}$$

Therefore the conditional posterior is obtained as a zero gradient point of the following Lagrangian

$$\begin{aligned} L_\pi = \; & G_\pi + V_\pi \\ & + \alpha_k\left(\mathbf{h}_k\right)\left[q\left(\mathbf{h}_k|\pi\right) - \sum_{\mathbf{h}_{k+1}} q\left(\mathbf{h}_{k+1},\mathbf{h}_k|\pi\right)\right] \\ & + \beta_k\left(\mathbf{h}_k\right)\left[q\left(\mathbf{h}_k|\pi\right) - \sum_{\mathbf{h}_{k-1}} q\left(\mathbf{h}_k,\mathbf{h}_{k-1}|\pi\right)\right] \\ & + \gamma_k\left(\mathbf{h}_k\right)\left[q\left(\mathbf{h}_k|\pi\right) - \sum_{\mathbf{o}_k} q\left(\mathbf{o}_k,\mathbf{h}_k|\pi\right)\right] \\ & + \delta_k\left(\mathbf{o}_k\right)\left[q\left(\mathbf{o}_k|\pi\right) - \sum_{\mathbf{h}_k} q\left(\mathbf{o}_k,\mathbf{h}_k|\pi\right)\right], \end{aligned}$$

where $\alpha_k, \beta_k, \gamma_k$, and δ_k denote the Lagrange multipliers for the corresponding equality constrain.

The *u*pdate equations for the conditional posterior are obtained as the zero gradient points of the Langrangian (Yedidia et al., 2001b, 2003a) defined above, therefore

$$q\left(\mathbf{o}_k,\mathbf{h}_k|\pi\right) = \frac{\bar{p}\left(\mathbf{o}_k\right) p\left(\mathbf{o}_k|\mathbf{h}_k\right) m_\pi^{k+1}\left(\mathbf{h}_k\right) m_\pi^{k-1}\left(\mathbf{h}_k\right)}{Z_k^\pi} \tag{2.30}$$

$$q\left(\mathbf{o}_k|\pi\right) = \frac{\bar{p}\left(\mathbf{o}_k\right) m_\pi^{k}\left(\mathbf{o}_k\right)}{Z_k^\pi} \tag{2.31}$$

$$\begin{aligned} q\left(\mathbf{h}_k,\mathbf{h}_{k-1}|\pi\right) = & \frac{p\left(\mathbf{h}_k|\mathbf{h}_{k-1},\pi\right)}{Z_{k,k-1}^\pi} \\ & \prod_{i=k-1}^{k} m^i\left(\mathbf{h}_i\right) m_\pi^{k+1}\left(\mathbf{h}_k\right) m_\pi^{k-2}\left(\mathbf{h}_{k-1}\right) \end{aligned} \tag{2.32}$$

$$q\left(\mathbf{h}_k|\pi\right) = \frac{m^k\left(\mathbf{h}_k\right) m_\pi^{k+1}\left(\mathbf{h}_k\right) m_\pi^{k-1}\left(\mathbf{h}_k\right)}{Z_k^\pi} \tag{2.33}$$

$$q\left(\pi\right) = \frac{p\left(\pi\right) \mathrm{e}^{-G_\pi - V_\pi}}{\sum_\rho p\left(\rho\right) \mathrm{e}^{-G_\rho - V_\rho}} \tag{2.34}$$

where $m_\pi^i\left(x_j\right)$ denotes a message from the ith node that is a direct neighbor to the jth node, for $x_j \in \{\mathbf{h}_k, \mathbf{o}_k\}$. Also, to simplify the notation we have used the following relation for $k \leq t$

$$\bar{p}\left(\mathbf{o}_k\right) = \begin{cases} 1, & \text{if } \mathbf{o}_k = \underline{\mathbf{o}}_k, \\ 0, & \text{otherwise} \end{cases}$$

The messages are computed iteratively as follows

$$
\begin{aligned}
m^{k}\left(\mathbf{h}_{k}\right) &= \sum_{\mathbf{o}_{k}} \bar{p}\left(\mathbf{o}_{k}\right) p\left(\mathbf{o}_{k}|\mathbf{h}_{k}\right), \\
m_{\pi}^{k}\left(\mathbf{o}_{k}\right) &= \sum_{\mathbf{h}_{k}} p\left(\mathbf{o}_{k}|\mathbf{h}_{\tau}\right) m_{\pi}^{k+1}\left(\mathbf{h}_{k}\right) m_{\pi}^{k-1}\left(\mathbf{h}_{k}\right), \\
m_{\pi}^{k+1}\left(\mathbf{h}_{k}\right) &= \frac{1}{Z_{k,\pi}^{\prime}} \sum_{\mathbf{h}_{k+1}} p\left(\mathbf{h}_{k+1}|\mathbf{h}_{k}\right) m^{k+1}\left(\mathbf{h}_{k+1}\right) m_{\pi}^{k+2}\left(\mathbf{h}_{k+1}\right), \\
m_{\pi}^{k-1}\left(\mathbf{h}_{k}\right) &= \frac{1}{Z_{k,\pi}^{\prime\prime}} \sum_{\mathbf{h}_{k-1}} p\left(\mathbf{h}_{k}|\mathbf{h}_{k-1}\right) m^{k-1}\left(\mathbf{h}_{k-1}\right) m_{\pi}^{k-2}\left(\mathbf{h}_{k-1}\right),
\end{aligned}
\tag{2.35}
$$

Figure 2.3 shows a graphical representation of the posterior beliefs and messages on the graph of the generative model. Information from forward and backward inference processes is integrated for perception and planning. We denote these distinct pathways as *f*orward messages and *b*ackward messages, respectively. Forward messages carry information from the past to the future, given the observations that were made and the states that were inferred. Backward messages pass back information from the prior beliefs about future outcomes and their corresponding states, and from observations already made to update the estimates of earlier states. The messages will be different for different control states, which makes them dependent on the policy π. For graphs without loops these update rules converge to a unique solution at the global minimum of the free energy, for which approximate marginals equal the marginals of the true posterior $q\left(x_{j}|\pi\right) = f\left(x_{j}|\underline{\mathbf{o}}, \pi\right)$ (Pearl, 1988; Yedidia et al., 2001b; Yedidia, Freeman, & Weiss, 2001a). Note, that the beliefs do not converge to the posterior $p(x_{j}|\underline{\mathbf{o}}, \pi)$ according to the generative model, but to the true posterior $f(x_{j}|\underline{\mathbf{o}}, \pi)$. This means that the beliefs do not correspond to optimal predictions, but are averaged over expected (preferred) future outcomes.

Combining the backwards and forwards messages corresponds to an evaluation of the variational free energy for each policy, so that an approximate posterior probability distribution for following a policy can be inferred. Inserting the update Equations (2.30) and (2.33) into the free energy Equation (2.28) yields the following relation for the conditional free energy

$$
\begin{aligned}
F_{\pi} &= G_{\pi} + V_{\pi} \\
&= -\ln Z_{T}^{\pi} - \sum_{k=1}^{T} \ln Z_{k,\pi}^{\prime\prime}
\end{aligned}
\tag{2.36}
$$

The posterior probability of following a policy π in accordance with the prior beliefs $\bar{p}\left(\tilde{\mathbf{o}}\right)$ is then obtained by inserting (2.36) into (2.34).

2.3.5 Action selection

To reach goals, the agent has to generate a sequence of actions, or in other words, the agent has to select a behavioral model which is most likely to fulfill its goals. One possible mechanism is to select the mode of the posterior over behavioral policies

$$
\begin{aligned}
\pi^{*} &= \arg\max q\left(\pi\right) \\
u_{t} &= \pi^{*}(t) = u_{t}^{*}
\end{aligned}
\tag{2.37}
$$

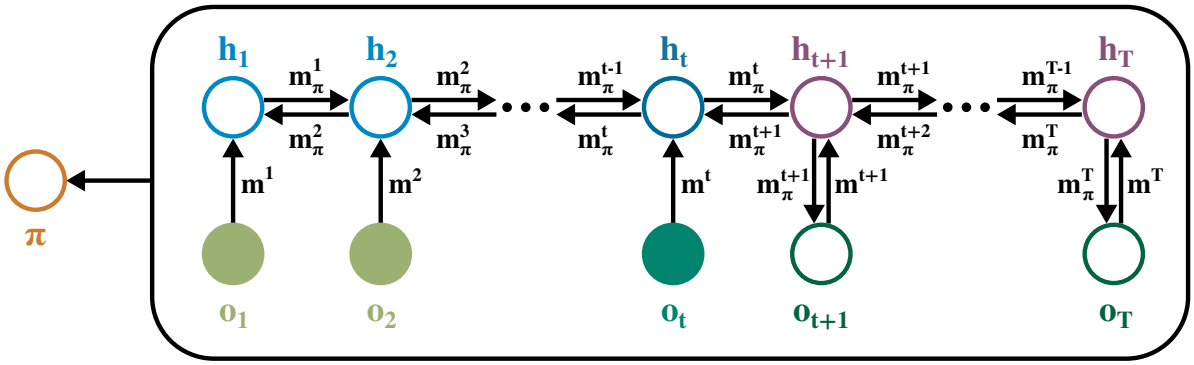

Figure 2.3: Graphical presentation of the model inversion under active inference
The notation used here corresponds to the one in Figure 2.2. However, the arrows now indicate the messages that are passed from one node to another. The arrows pointing up are the messages $m^k(\mathbf{h}_k)$ from an observation to the respective state, which influence the inference which state had been visited or should be visited in the future. The arrows pointing right correspond to the forward messages $m_\pi^{k-1}(\mathbf{h}_k)$. The arrows pointing to the left represent the backward messages $m_\pi^{k+1}(\mathbf{h}_k)$. The arrows pointing down are the messages $m_\pi^k(\mathbf{o}_k)$ from an estimated future state to their corresponding observation. They shape the estimate of what will be observed in the future. Note that the last three described messages depend on the policy π, i.e. they will be different for each sequence of control states. In that manner, they influence the estimate of the policy π, and thereby determine the probability of following a policy (arrow pointing from the big box to the policy).

and select the respective action u_t^* at time step t. We will call this type of action selection *m*aximum selection.

Another approach is model averaging in which the agent uses its posterior beliefs over policies to build expectations over control states

$$q\left(u_t|\underline{u}\right) = \sum_\pi p\left(u_t|\underline{u},\pi\right) q\left(\pi\right), \tag{2.38}$$

where the chosen action is sampled from $q\left(u_t|\underline{u}\right)$. We refer to this mechanism of action selection as *a*veraged selection.

For simplicity we will consider action selection to these two limiting cases. Note that it would be straightforward to introduce additional hidden variables which allow the agent to balance its behavior between model selection and model averaging (FitzGerald, Dolan, & Friston, 2014b), as previously proposed in (K. Friston et al., 2013).

2.3.6 Toy environment

To illustrate and compare the goal-directed behavior that results from the above derived update equations based on the mean-field and Bethe approximation, we will use a navigation task in a 4×4 grid world. The agent's task is to navigate from a starting position (red shaded square) to a goal position (blue shaded square), see Figure 2.4. Although a simple task, it is complex enough to illustrate the differences between the two approximations and providing insights into the limitations of the mean-field approximation.

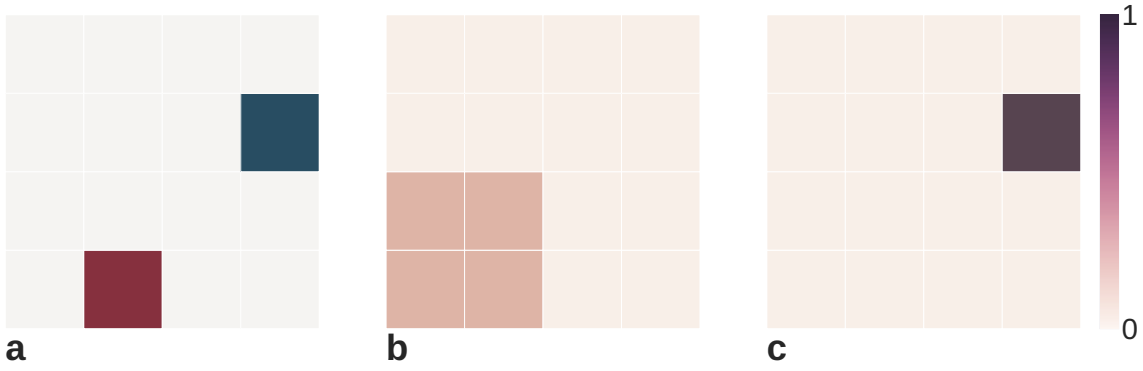

Figure 2.4: Grid world
a The agent starts out in the red shaded location and has to navigate to the blue shaded location on the grid world. **b** Prior beliefs over the starting state $p(\mathbf{h}_1)$ (color coded). **c** Prior beliefs over future outcomes $\bar{p}(\mathbf{o}_\tau)$ (color coded).

At each time step the agent makes an observation $\mathbf{o}_t$ that provides information about its current hidden state. In each state (node of the grid world) the agent can choose between $n_u = 4$ control states: go up, go down, go left, and go right. The task for the agent is to reach the goal state after making four choices. The number of time steps modelled in each run is $T = 5$. Note that if the agent is at a boundary, the movement into the direction of the boundary fails and the agent will not change its position.

After making an observation, the agent has to infer current and past states, and build expectations about future states and observations. This process corresponds to calculating the approximate posterior $q\left(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}|\pi\right)$. Given the policy-dependent posterior, the agent evaluates the total free energy F_π over all $n_\pi = 256$ possible policies. The total free energy defines the posterior beliefs over behavioral policies $q(\pi)$. In this specific environment, only six policies will lead to the goal state in the given time frame.

Before making any observations, the agent's beliefs are defined by its prior beliefs about his starting state $p(\mathbf{h}_1)$. To make the agent rely on observations when planning behavior, we let the agent be uncertain about its starting position by setting the prior beliefs to a uniform distribution over the four states in the bottom left corner (Figure 2.4b). To induce goal-directed behavior we have defined the prior beliefs over future outcomes $\bar{p}(\mathbf{o}_\tau)$ as a step function

$$\bar{p}(\mathbf{o}_\tau) = \begin{cases} \rho & \mathbf{o}_\tau = g \\ 1-\rho & \mathbf{o}_\tau \neq g \end{cases} \tag{2.39}$$

with constant values ρ for the goal observation g and $1-\rho$ for all other observations (Figure 2.4c). For simplicity we will consider the prior beliefs over future outcomes to be fixed to the same step function in all future time steps $t < \tau \leq T$ (effectively, our predicted free energy then accommodates a path integral of prior preferences).

To illustrate the agent's behavior, we will expose the agent to two different environments: (i) a grid world with varying observation uncertainty (Figure 2.5a) and (ii) a grid world with varying state transition uncertainty (Figure 2.5b). With increasing observation uncertainty the probability of making an observation associated with a neighboring state increases, while with increasing state transition uncertainty the probability of remaining in the current state increases.

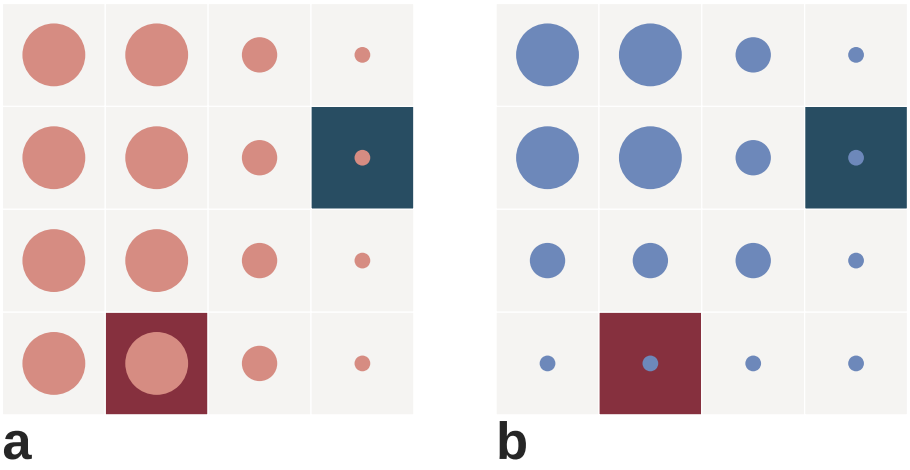

Figure 2.5: Experimental conditions
a The grid world with a varying observation uncertainty. The size of the circles scales with increasing observation uncertainty and decreases with a horizontal gradient from left to right. The agent starts out in a high uncertainty state, and it has to rely on inference about states to navigate through the grid. **b** The grid world with a varying state transition uncertainty. The size of the circles scales with increasing state transition uncertainty and decreases along a diagonal gradient from the upper left to the bottom and to the right. The positions in the bottom row and right-most column have no state transition uncertainty and state transitions from these positions are deterministic.

In the environment with varying observation uncertainty we have chosen a horizontal gradient of uncertainty, thus we defined the state dependent observation likelihood as

$$p\left(\mathbf{o}_k = i|\mathbf{h}_k = j\right) = \begin{cases} a_j & i = j \\ \frac{1-a_j}{n_j} & i \in N(j) \\ 0 & \text{otherwise} \end{cases} \tag{2.40}$$

where $a_j \in \left\{1, \frac{2}{3}, \frac{1}{2}, \frac{1}{2}\right\}$, $N(j)$ denotes the nearest neighbors of the jth node, and n_j the total number of neighbors of the jth node. To ensure that the goal state is associated with a single observation, we excluded it for the uncertainty specification from the neighborhood of all neighboring states. The specific value of a_j is inversely proportional to the size of the circle shown in Figure 2.4b. Note that the number of different observations corresponds to the number of different states hence $i, j \in \{1, \ldots, 16\}$.

In this environment, to make the agent rely on the inference about the state space in order to reach the goal, we set the initial state in the area with high observation uncertainty. Therefore, in the initial state and depending on the initial observation, the agent's beliefs will be distributed over the possible starting states. Whether the agent reaches the goal state or not strongly depends on the initial observation. Importantly, out of the policies which lead to the goal, some lead through the states with high observation uncertainty, while others lead to states with low observation uncertainty. An interesting question here is whether the agent follows more often policies that lead toward low observation uncertainty states, that is, whether the agent tends to reduce its initial uncertainty about the state space.

In the environment with state transition uncertainty, we removed the observation uncertainty, but chosen actions have a state dependent chance of failing. We have defined the state dependent transitioning probability as follows

$$p\left(\mathbf{h}_k = i|\mathbf{h}_{k-1} = j, u_t = a\right) = \begin{cases} b_j & j(a) = i \\ 1 - b_j & i = j, \\ 0 & \text{otherwise} \end{cases} \tag{2.41}$$

where $b_j \in \left\{1, \frac{2}{3}, \frac{1}{2}, \frac{1}{2}\right\}$, and $j(a)$ denotes the neighbor of node j in the direction of action a. If $j(a)$ points to the boundary then $b_j = 0$ for all boundary states j. As before the specific value of b_j is inversely proportional to the size of the circle in Figure 2.4a.

There is exactly one policy which leads to the goal state with certainty. We will consider this policy to be the optimal policy in this condition

$$\pi_{\text{optimal}} = (\text{right, right, up, up}) \tag{2.42}$$

2.4 Results

Here we present the behavioral differences between the Bethe approximation based agent and the mean-field approximation based agent for the two environments in the grid world. All presented cases were obtained as an average over 1,000 runs in each environment.

Figure 2.6: Success rates as a function of the magnitude of the prior beliefs over the goal observation ρ

a and **c** show the success rates for the observation uncertainty condition, **b** and **d** show the success rates for the state transition uncertainty condition. The top row (**a**,**b**) and the bottom row (**c**,**d**) show the results for averaged action selection and maximum action selection, respectively. The success rates of the Bethe agent are plotted in blue and the success rates of the mean-field agent in green, the transparent areas display the confidence intervals of 95%.

2.4.1 Prior preferences and performance

A model parameter with a strong influence on the agent's behavior is the prior over future outcomes $\bar{p}(\tilde{\mathbf{o}})$ (see Equation (2.39)). This prior defines the agent's preferences over future observations and modulates the predicted free energy of a behavioral policy (see Equation (2.19)). To investigate the impact of the prior preferences on the performance of the agents, we varied the value of the prior $\bar{p}(\mathbf{o}_\tau = g) = \rho$ between 0.5 and 0.999 and estimated the corresponding average success rate; defined as the percentage of trials in which the agent is at the goal location at the last time step T.

In Figure 2.6 we show the resulting success rates as a function of prior preference ρ in different conditions and action selection methods. Several patterns are clearly visible: (i) the success rates of agents using averaged action selection (top row of Figure 2.6) increase strongly with an increasing ρ, while the success rates of agents using maximum selection (bottom row) remain mostly constant and at higher levels compared to averaged selection; (ii) In the environment with observation uncertainty (left column) the Bethe agent achieves consistently higher success rates, independent of the action selection method; (iii) in the environment with state transition uncertainty the success rates of the agents are closely matched, with a slight advantage of the mean-field agent using the averaged selection for

high prior preferences. In what follows we will explain what gives rise to this specific pattern of performance differences between agents and action selection methods.

The influence of the prior preferences ρ on the success rates depends on the components that define posterior beliefs over policies. The key factor which determines the value of the conditional free energy F_π—and therewith the posterior $q(\pi)$—is the cross entropy $-\sum_{\mathbf{o}_\tau} q\left(\mathbf{o}_\tau|\pi\right) \ln \bar{p}\left(\mathbf{o}_\tau\right)$. Hence, the ranking of the policies is independent on the value of prior preference ρ, however their relative probabilities change. In other words, in the case of maximum selection the value of ρ does not influence which policy is selected by the agent, whereas in the case of averaged selection the relative value of different policies has an effect on action selection. Thus an increasing ρ under averaged selection makes the agents' behavior more goal-directed, and thereby more successful.

2.4.2 Prediction accuracy

To pinpoint the reason for the large difference in the performance between the two agents in the environment with observation uncertainty, we looked into the posterior beliefs over policies evaluated in the first time step $t = 1$. Because the predicted probability of making the preferred observation in the final time step $q\left(\mathbf{o}_T = g|\pi\right)$ is the main contributor to the probability $q(\pi)$ of following a policy, we examined if the agents correctly predict that they will or will not reach the goal state when evaluating policies.

To do this, we calculated the true positive and false positive classification rates. When an agent correctly predicted reaching the goal state when evaluating one of the 6 policies which lead to the goal, we counted this as a true positive. When the agent incorrectly predicted reaching the goal when evaluating one of the remaining policies, we counted this as a false positive. Figure 2.7a shows the true positive classification rate of both agents in the first time step. The Bethe agent has a 95% true positive rate, meaning that, when evaluating a policy that could lead to a goal, it almost always correctly predicts that the policy will be successful. In contrast, the mean-field agent has a true positive rate below 60%, incorrectly classifying policies as not leading to the goal state, despite them being successful policies. This low true positive rate skews the approximate posterior $q(\pi)$, so that policies which would be good to follow have a low value, leading to erroneous behavior, and explaining the second effect of the overall lower success rates.

In Figure 2.7b the false positive values are shown. The Bethe agent has a false positive rate close to 0%, whereas the mean-field agent has always a false positive rate greater than zero. In other words, the Bethe agent almost never assigns non zero values to policies that do not lead to the goal state, whereas the mean-field agent predicts that some policies will lead to the goal, when they do not. The false positive rate of the mean-field agent increases with an increasing value of the prior preference over goal outcome ρ. This gives rise to the third effect, the drop in success rates for high prior values ρ, as the agent will follow policies which can not lead to the goal state.

These differences in performance of the two agents can be related to the sensitivity of the gradient descent procedure (see Equation (2.25)) to the initial conditions (see Equation (2.26)). Indeed, we observe that changing the initial conditions of the gradient descent influences the final solutions, and hence the performance of the mean field agent. However, for different environments, different initial conditions are required to improve the performance of the mean-field agent. This points to an underlying issue of the mean-field approximation when applied to sequential inference: We found that the approximate posterior over future states

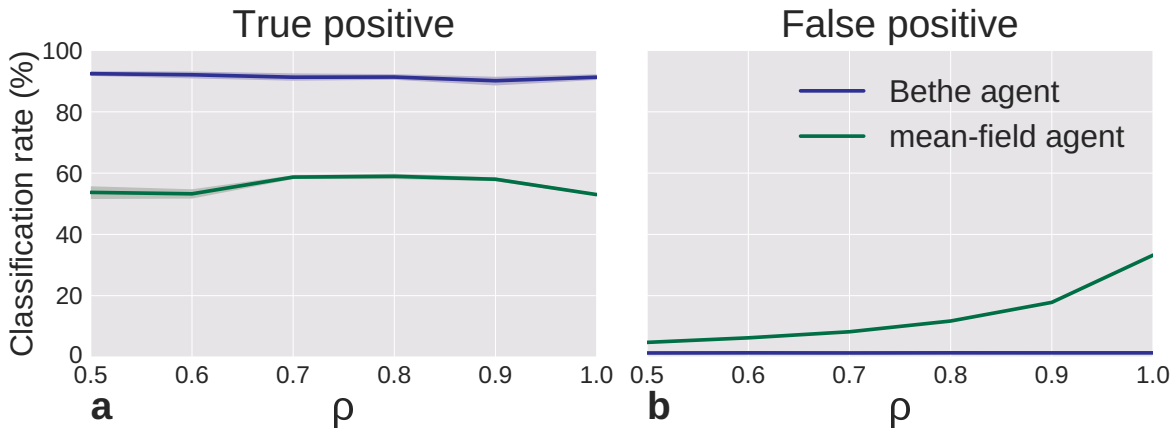

Figure 2.7: Classification of policies by the agents in the first time step $t = 1$
True and false positive classification rates of policies in the environment with observation uncertainty for different values of the prior over the goal observation ρ. a percentage of policies correctly classified as leading to the goal by the Bethe agent (blue) and mean-field agent (green), out of policies which would lead to the goal in a deterministic environment (true positives). b percentage of policies incorrectly classified to be leading to the goal by the agent, out of policies which do not lead to the goal (false positives).

can converge to impossible state space configurations. Thus, the agent predicts that it will execute an impossible state transition, i.e. jump across the grid, which causes an erroneous evaluation of the posterior over policies and elicits unfavorable behavior.

Interestingly, the closely matched success rates and the higher success rate of the mean-field agent in the environment with state transition uncertainty can also be related to the prediction of impossible state transitions. Even in the environment with state transition uncertainty, the mean-field agent accurately predicts the goal state only for the optimal policy (the path without transition uncertainty). For other policies we again observe predictions of impossible state transitions for the majority of policies. This erroneous inference leads to an higher posterior value of the optimal policy $q\left(\pi_{\text{optimal}}\right)$, which in effect improves the mean-field agent's performance, as it results in higher probability of following optimal policy when using averaged selection (Figure 2.6b). Importantly, the higher the ρ is the larger is the penalty for policies predicted not to reach the goal state, which makes the mean-field agent better than the Bethe agent for the largest ρ.

2.4.3 Optimal policy selection

To illustrate the differences in agents' behavior in the two environments, we show in Figure 2.8 and Figure 2.9 the average paths followed by the agent for the prior preference fixed to $\rho = 0.999$.

In the case of the environment with observation uncertainty (see Figure 2.8) we see clear differences between the selected paths of the Bethe and mean-field agents. In contrast to the mean-field agent, the Bethe agents consistently follows only goal reaching policies in a fairly symmetric selected path structure. The slight bias towards policies going to the right is not a result of the agents' higher valuation of policies that reduce uncertainty about the state space, but is due to the stochastic nature of the first observation and the subsequent

difference in inference about the starting state. Indeed, we find that the initial uncertainty about the occupied state is passed on to predictions about future states, so that the entropy of the agents' estimate about future states does not decrease, even when evaluating a policy which contains an informative (low uncertainty) state (see Section 2.5).

Although the mean-field agent follows similar paths when reaching the goal state, it surprisingly selects policies which lead to the left, away from the goal. These are stunning examples of trajectories where the agent falsely predicts that some policies will lead to the goal when they do not. The cause of this behavior is erroneous inference about the initial state in the presence of observation uncertainty, leading to false beliefs that the goal is not reachable from its initial state. When the agent believes that it is too far from the goal state, all policies are treated as equally likely, as the expectation is that none of them would lead to the goal state. This is why the agent chooses steps to the left even in maximum selection mode.

Interestingly, the false predictions of the mean-field agent (the convergence of posterior beliefs to impossible trajectories) are the main factor driving the behavior in the environment with observation uncertainty. Here, the agent's overconfidence about current policies and current states prevents it from switching to a different policy, even though the observations do not carry sufficient information. Furthermore, the mean-field agent shows a strong preference for policies leading through the high uncertainty regime under maximum selection. We found the reason for this lies in reduced convergence issues for beliefs over future states (more accurate representation of state transition paths) when policies that lead through high uncertainty regions are evaluated. The true positive rate for these policies is higher than for other policies leading to the goal.

In the environment with state transition uncertainty (Figure 2.9), the behavior of the two agents is very similar. Importantly, in the case of maximum action selection both agents correctly valued the path with least uncertainty as optimal, hence they always choose the optimal policy. In averaged selection mode, when actions are chosen by averaging over the values of policies, non-optimal actions have a non-zero probability of being chosen. This effect increases with the number of policies. This causes a branching out from the optimal path and a subsequent drop in success rate. As discussed above, avoiding uncertainty is not a driving factor in the agent's evaluation of policies. Rather, policies are weighted according to the probability of reaching the goal state.

In summary, we found severe drawbacks in the mean-field agent's planning process. When inferring the future states for a given policy, the agent's beliefs would converge to impossible configurations of future states. In our formulation, both forward and backward messages shape the beliefs about the future. Such a setup leads to multi-modal true posteriors as a result of a divergence between the forward and backward predictions. Under the gradient descend procedure used here (see Equation (2.25)) for the mean-field agent, its beliefs settle around one of the modes (local optimum of the free energy). Since the value and probability of a policy are determined by the predicted probability of reaching the goal state, inaccurate beliefs about future states lead to inaccurate posterior beliefs over possible policies. Depending on the environment, this anomalous inference can lead to either reduced or increased performance of the mean-field agent.

The Bethe agent, however, in our simulations, always accurately predicted future states given past observations, as the pairwise statistical dependencies explicitly prevent a divergence. Furthermore, the beliefs were able to better maintain this multi-modality stemming from the superposition of the forward and backward messages. And as the convergence of the beliefs to the true posterior is guaranteed under the belief propagation update rules, the Bethe agent will always optimally predict future states. A correct prediction of the probability of

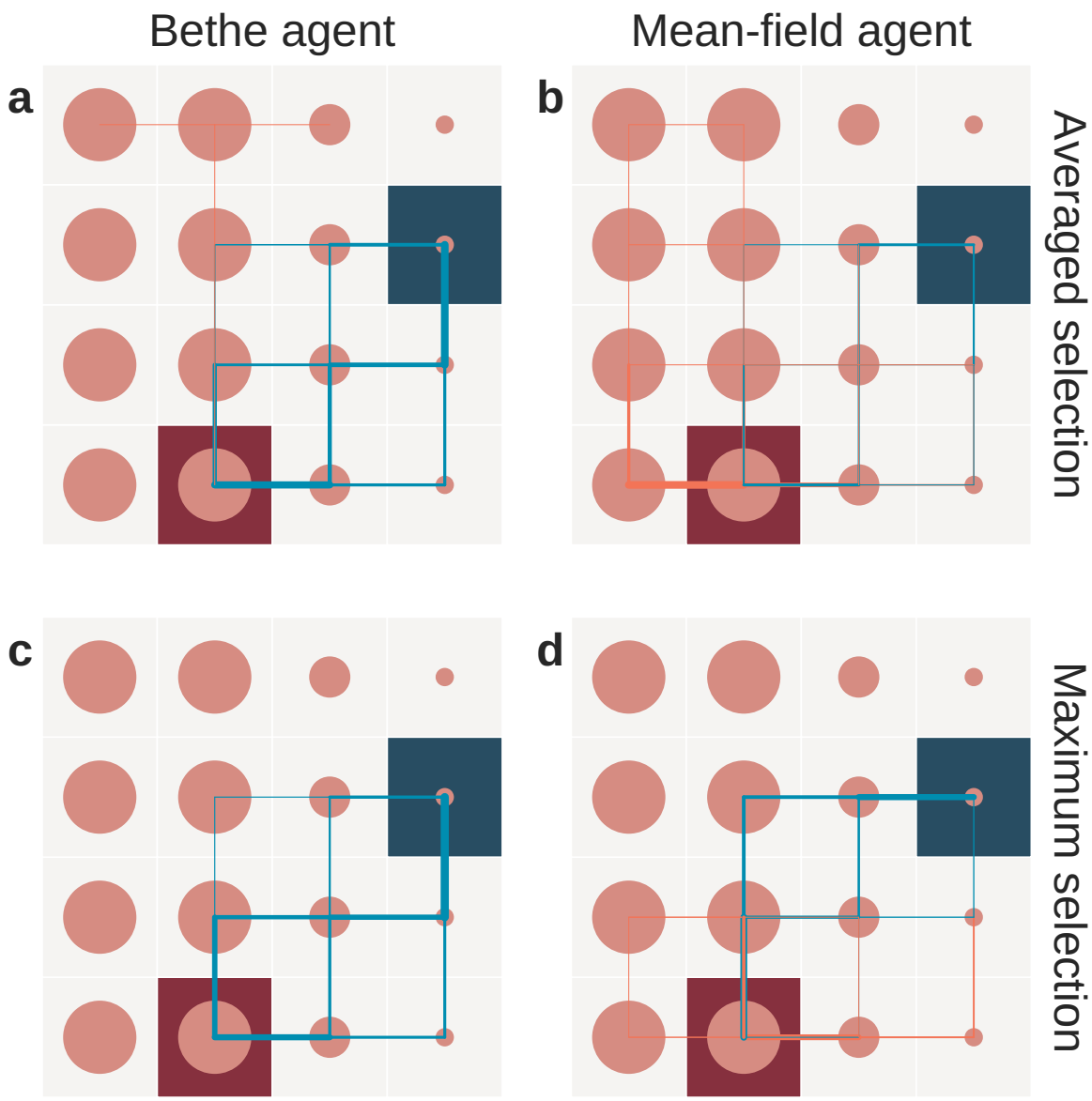

Figure 2.8: Simulation results in the environment with observation uncertainty
The left column (a, c) shows the trajectories for the Bethe agent and the right column (b, d) the results for the mean-field agent. The cyan lines indicate the paths chosen by the agent in successful runs. The red lines indicate paths chosen by the agents in unsuccessful runs. Their thickness reflects the frequency with which a certain path was followed.

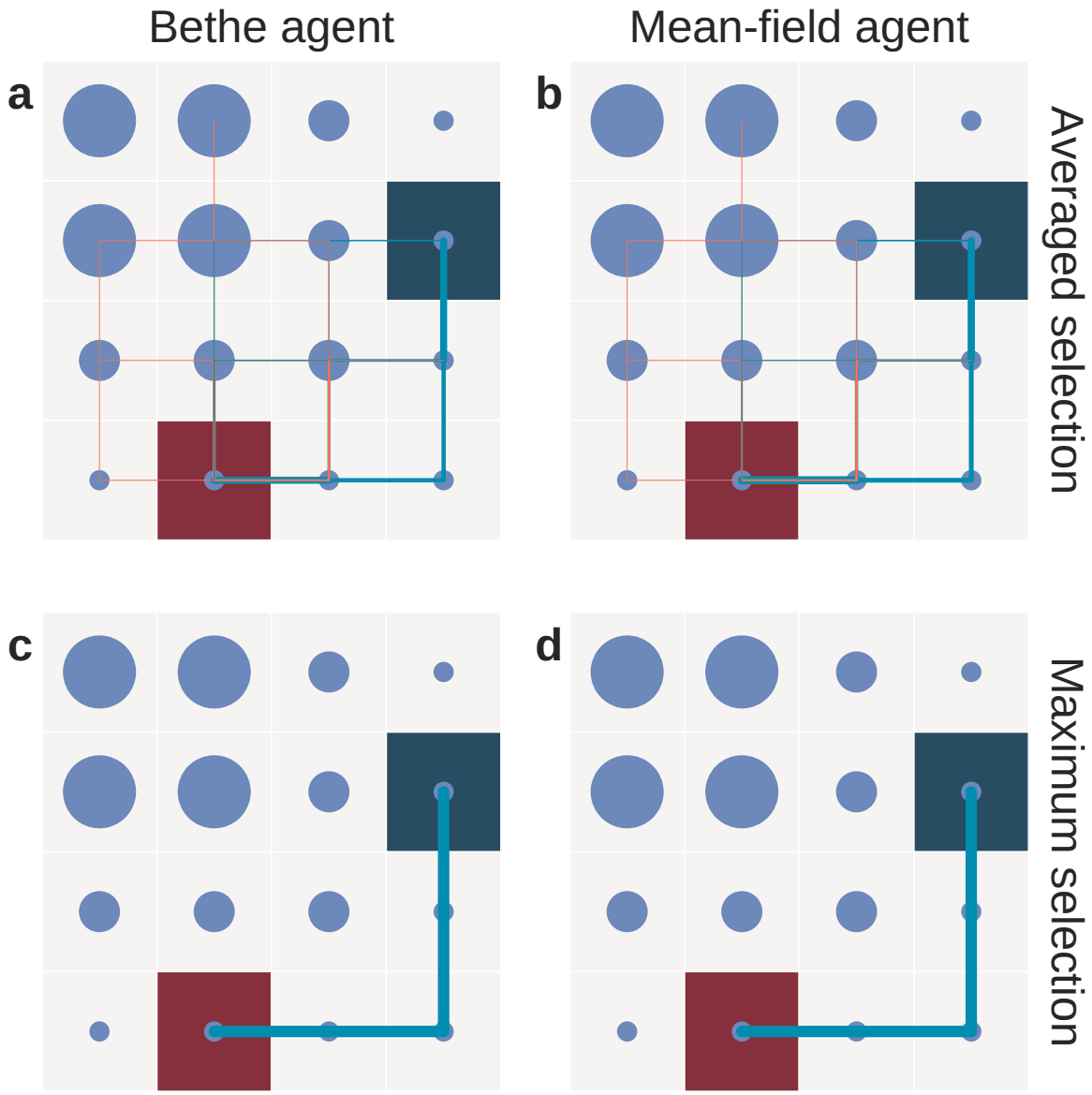

Figure 2.9: Simulation results in the environment with state transition uncertainty
The left column (**a**, **c**) shows the trajectories for the Bethe agent and the right column (**b**, **d**) the results for the mean-field agent. The two rows show the results for the two ways of action selection and the paths are color coded as in Figure 2.8.

reaching the goal state automatically leads to a more accurate policy evaluation as compared to the mean-field agent.

2.5 Discussion

We revisited a specific solution of planning as inference for modelling goal-directed behavior given by the active inference framework, where posterior beliefs about hidden states, future observations and policies are obtained by minimizing the variational free energy. Importantly, we provide here an alternative approach to the derivation of the key update equations of active inference agents. In contrast to previous formulations of active inference, the agent's behavior aims at minimizing the expectation over the predicted free energy, instead of the expected free energy as postulated previously (K. Friston et al., 2015; K. Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty, & Pezzulo, 2016; K. Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016). This allowed us to reveal the effects of the mean field approximation in the face of uncertainty. In future work, we will investigate and compare behavior that results from both formulations.

Besides the typically used mean-field approximation (K. Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty, & Pezzulo, 2016; K. Friston et al., 2015) we provide a variational treatment of planning as inference based on the Bethe approximation. In contrast to the mean-field approximation—under which statistical independence of hidden variables is assumed—the Bethe approximation assumes pairwise statistical dependencies between hidden variables in the approximate posterior. To demonstrate the key differences between acting agents based on the Bethe approximation and the mean-field approximation we have designed two illustrative toy environments in which the agents had to perform a multi-trial goal-reaching task while being exposed to either observation uncertainty or state transition uncertainty. We found that assuming pairwise statistical dependence between hidden variables improves an agent's inference of hidden states. This leads to more accurate predictions about the future, and consequently, evaluation of policies. These improvements resulted in more optimal goal-directed behavior and higher success rates.

In the environment with observation uncertainty (Figure 2.5a), the state estimation was dependent on noisy observations. This environment illustrates a condition in which goal-directed behavior is generated under limited information about the current state of the environment. For example, in a maze task an agent might not exactly know where it is, due to ambiguity in the environment. Here, the Bethe agent showed consistently and dramatically higher success rates in goal-reaching behavior due to a more robust, policy-dependent inference of past, current, and future states and observations. We linked the low success rates of the mean-field agent to the erroneous formation of beliefs about hidden states. This misrepresentation of hidden states is caused by the convergence of posterior beliefs to configurations that are impossible under any given policy. This is due to the fact that agents infer the sequence of most probable states rather than the most probable sequence. When dealing with inference under uncertainty, the true posterior is often a multi-modal distribution. However, under the gradient descent procedure used here, the posterior beliefs mostly converged to uni-modal distributions so that one of the peaks of the true multi-modal distribution becomes enlarged, while all other peaks vanish. As a result, the agent either misrepresents uncertainties, so that its beliefs only represent the most likely state, or the agent predicts states which are impossible from the perspective of forward planning, but are likely from the perspective of backward planning, i.e. going from the goal state backwards.

Due to the over-confidence in beliefs over current states and expectations about future states the mean-field agent cannot recover from an initial erroneous inference. This even holds after sampling more observations and forming more accurate beliefs over hidden states. This in turn leads to an erroneous evaluation of behavioral policies. In contrast, the Bethe agent was able to rapidly adjust its evaluation of policies, even if it had been misled by a first, noisy observation.

Although a possible remedy for the mean-field agent may be to adapt the initial conditions in the gradient descent optimization procedure, these initial conditions would most likely be, as we found for our simulations, environment- and task-specific. Another way to resolve this issue for the mean-field agent might be to use a more sophisticated method than a simple gradient descent. It would also be possible to base the predictions only on the forward inference process (as done in previous work (K. Friston et al., 2015; K. Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty, & Pezzulo, 2016; K. Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016)), instead of combining forward and backward inference. While this would lead to more accurate predictions of the future, and possibly less convergence issues, it would strip the agent of the possibility to infer which states are on its way to the goal. We found that the Bethe approximation provides a principled solution, as it is able to capture the temporal structure of the environment and convergence to global optima is typically guaranteed in a sequential decision task environment.

In the environment with state transition uncertainty (Figure 2.5b), hidden states were directly observable, but actions were executed stochastically. Here, the effect of erroneous state space representation on success rates of the mean-field agent was reduced, in comparison to the environment with observation uncertainty. Both agents avoided high uncertainty regions, illustrating that the driving factor in goal-directed behavior is the predicted probability of reaching the goal state.

Such avoidance of high uncertainty states was not seen in the observation uncertainty condition, showing that agents do not intrinsically value informative states in our formulation using the predictive free energy, in contrast to previous formulations of active inference (K. Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016; Schwartenbeck et al., 2015). Visiting a state associated with low observation uncertainty can be interpreted as information gathering, as the observation would be more informative about the underlying hidden state. We did not observe this behavior in the agents, which we relate to the fact that initial uncertainty about the state space is passed on to predictions about future states, keeping the expected entropy of a future state high and thereby making such a state not more valuable to the agent. In previous work on active inference (K. Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016; Schwartenbeck et al., 2015), policy evaluation was done using a prior over policies which was defined using the expected free energy. The expected free energy contains a term evaluating the epistemic value (the informativeness of an action) of each policy. Using the expected free energy, agents follow informative policies with high epistemic value, meaning they tend to visit states with low observation uncertainty. As the epistemic value term does not follow under the derivation presented here (see Section 2.3.4), where we derived a policy evaluation based on the predicted free energy, it is not surprising that we do not observe such behavior (see Appendix for details on the expected free energy).

The formulation of active inference under the mean-field ansatz has previously been put forward as a process theory of neuronal function (K. Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016). Furthermore, (K. J. Friston, Parr, & de Vries, 2017) recently proposed a neuronal connection scheme for belief propagation update rules under active inference. However, the authors considered a modified belief propagation scheme in which the condi-

tional dependencies among hidden states are ignored, hence allowing them to obtain update rules using the mean-field approximation. Under the Bethe approximation, the interpretation in terms of neural coding does not necessarily change, and can be linked to past work on possible implementations of belief propagation in neuronal networks. For example, (Shon & Rao, 2005; Ott & Stoop, 2007) demonstrated an implementation of belief propagation using a neuronal network in cases when the generative model contains only pairwise interactions (like Bayesian graphs or Markov random fields). In this formulation, neurons are interpreted as nodes of the graph of the generative model, and connections as conditional probabilities. In this scheme, the intuitive idea is that the activation of neurons encodes the beliefs about hidden variables, while the messages are transmitted by neural signal transaction. Similarly, (Deneve, 2005) showed as a proof of principle that inference based on belief propagation can be implemented in a network of spiking neurons. Interestingly, following this line of work, (T. S. Lee & Mumford, 2003b; Jardri & Denève, 2013) discussed a possible link between belief propagation in cortical networks and optical illusions and hallucinations.

A potential issue with neuronal implementation of belief propagation arises when the generative model becomes more complex than the one used in this work. For example, it might require interaction of more than two variables. Mathematically, the Bethe approximation and the resulting belief propagation update equations scale well to these more complex models. However, in this case, the mapping of conditional beliefs and messages to neuronal architecture becomes more challenging and is subject to ongoing discussion. It might be necessary to have an extra neuronal pool to calculate the messages (George & Hawkins, 2009; Steimer, Maass, & Douglas, 2009).

An example of a more complex model is a hierarchical generative model (K. J. Friston, Rosch, Parr, Price, & Bowman, 2017). Here, a mixture of approximate representations of the posterior could be used. In this case, different levels of the hierarchy could be represented independently in the posterior (mean-field approximation) and pairwise interactions would only be captured within the same levels of representation (Bethe approximation). Additionally, learning principles have recently been introduced to active inference (K. Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty, & Pezzulo, 2016; K. Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016) which could easily be combined with the Bethe approximation. It would be interesting in the future to explore whether the appropriate factorization of the posterior can be learned over time, which could lead to an emergence of the most effective approximation of a task environment.

In summary, we have presented a method for incorporating belief propagation within the active inference framework using the Bethe approximation. The presented update equations of the active inference framework complement past work (K. Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016; K. Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty, & Pezzulo, 2016; K. Friston et al., 2015) and extend, in principle, the application range of active inference to complex behavioral tasks with various sources of uncertainty.

2.6 Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (SFB 940/2, Projects A9 and Z2).

2.7 Appendix

2.7.1 Relation between the predicted and expected free energy

In contrast to the variational free energy (which is a functional of a distribution over hidden states and future observations, given observed outcomes) the expected free energy can be expressed as the expectation over future (unobserved) outcomes, given a policy that defines future beliefs over states (Kaplan & Friston, 2017). Alternatively we can express the expected free energy as

$$\begin{aligned} G_{\pi}^{\text{expected}} &= \sum_{\tilde{\mathbf{o}},\tilde{\mathbf{h}}} p\left(\tilde{\mathbf{o}},\tilde{\mathbf{h}}|\pi\right)\left[\ln p\left(\tilde{\mathbf{h}}|\pi\right) - \ln p\left(\tilde{\mathbf{h}}|\tilde{\mathbf{o}},\pi\right) - \ln \bar{p}\left(\tilde{\mathbf{o}}\right)\right] \\ &= -\sum_{\tilde{o},\tilde{\mathbf{h}}} p\left(\tilde{\mathbf{o}},\tilde{\mathbf{h}}|\pi\right)\left[\ln \frac{p\left(\tilde{\mathbf{h}}|\tilde{\mathbf{o}},\pi\right)}{p\left(\tilde{\mathbf{h}}|\pi\right)} + \ln \bar{p}\left(\tilde{\mathbf{o}}\right)\right] \\ &= \sum_{\tilde{o},\tilde{\mathbf{h}}} p\left(\tilde{\mathbf{o}},\tilde{\mathbf{h}}|\pi\right)\left[\ln \frac{p\left(\tilde{\mathbf{o}}|\pi\right)}{\bar{p}\left(\tilde{\mathbf{o}}\right)} - \ln p\left(\tilde{\mathbf{o}}|\tilde{\mathbf{h}}\right)\right] \end{aligned} \tag{2.43}$$

As an agent maintains only approximate estimates of the beliefs over future states and outcomes we obtain the approximate form of the expected free energy for $p\left(\tilde{\mathbf{o}},\tilde{\mathbf{h}}|\pi\right) \approx q\left(\tilde{\mathbf{o}},\tilde{\mathbf{h}}|\pi\right)$.

Under the mean-field approximation (see Equation (2.20)) the expected free energy at future time step τ becomes

$$G_{\pi}^{\text{expected}}\left(\tau\right) = D_{KL}\left[q\left(\mathbf{o}_{\tau}|\pi\right)||\bar{p}\left(\mathbf{o}_{\tau}\right)\right] + \sum_{\mathbf{h}_{\tau}} q\left(\mathbf{h}_{\tau}|\pi\right) H\left[p\left(\mathbf{o}_{\tau}|\mathbf{h}_{\tau}\right)\right]. \tag{2.44}$$

In contrast, under the Bethe approximation the expected free energy becomes

$$G_{\pi}^{\text{expected}}\left(\tau\right) = \sum_{\mathbf{o}_{\tau:t+1}} q\left(\mathbf{o}_{\tau:t+1}|\pi\right) \ln \frac{q\left(\mathbf{o}_{\tau}|\mathbf{o}_{\tau-1:t+1}\pi\right)}{\bar{p}\left(\mathbf{o}_{\tau}\right)} + \sum_{\mathbf{h}_{\tau}} q\left(\mathbf{h}_{\tau}|\pi\right) H\left[p\left(\mathbf{o}_{\tau}|\mathbf{h}_{\tau}\right)\right], \tag{2.45}$$

as under the Bethe approximation the beliefs over future outcomes do not factorize into the product over marginals at each time step.

In this formulation, the expected free energy contains two terms: The first term encodes the extrinsic value of a policy, as it is minimized when the agent predicts that a specific policy will fulfill the prior expectations over future outcomes. The second term defines the expected ambiguity, that is, expected observational uncertainty at future time steps τ. This term is minimized when an agent visits informative states.

The expected free energy expresses a slightly different set of terms compared to the predicted free energy. To show the similarities and differences, we can rewrite the predicted

free energy (Equation (2.19)) as

$$
\begin{aligned}
G[q] = & \sum_{\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi} q\left(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi\right) \left[\ln \frac{q\left(\tilde{\mathbf{o}} | \underline{\mathbf{h}}, \pi\right)}{\bar{p}\left(\tilde{\mathbf{o}}\right)} - \ln p\left(\tilde{\mathbf{o}} | \tilde{\mathbf{h}}\right)\right] \\
& + \sum_{\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi} q\left(\tilde{\mathbf{o}}, \mathbf{h}_{1:T}, \pi\right) \ln \frac{q\left(\tilde{\mathbf{h}} | \tilde{\mathbf{o}}, \underline{\mathbf{h}}, \pi\right)}{p\left(\tilde{\mathbf{h}} | \mathbf{h}_t, \pi\right)}.
\end{aligned}
\tag{2.46}
$$

In this form, the predicted free energy is similar to the expected free energy (Equation (2.43)), and contains two pragmatic terms and a term similar in form to the information gain of the expected free energy, albeit with an opposite sign. The expected free energy can be recovered from the predicted free energy by imposing the constraint $\sum q\left(\tilde{\mathbf{o}}, \mathbf{h}\right) 1 : t, \pi) \ln q\left(\tilde{\mathbf{h}} | \tilde{\mathbf{o}}, \underline{h}, \pi\right) = \sum q\left(\tilde{\mathbf{o}}, \mathbf{h}\right) 1 : t, \pi) \, p\left(\tilde{\mathbf{h}} | \mathbf{h}_t, \pi\right)$. The interpretation of the third term becomes more obvious when the respective approximations are inserted into the predicted free energy.

Under the mean-field approximation (Equation (2.22)), the predicted free energy can be rearranged as

$$
\begin{aligned}
G_\pi\left(\tau\right) = & \sum_{\mathbf{o}_\tau, \mathbf{h}_\tau} q\left(\mathbf{o}_\tau, \mathbf{h}_\tau | \pi\right) \left[\ln \frac{q\left(\mathbf{o}_\tau | \pi\right)}{\bar{p}\left(\mathbf{o}_\tau\right)} - \ln p\left(\mathbf{o}_\tau | \mathbf{h}_\tau\right)\right] \\
& + \sum_{\mathbf{o}_\tau, \mathbf{h}_\tau, \mathbf{h}_{\tau-1}} q\left(\mathbf{o}_\tau, \mathbf{h}_\tau | \pi\right) q\left(\mathbf{h}_{\tau-1} | \pi\right) \ln \frac{q\left(\mathbf{h}_\tau | \mathbf{o}_\tau, \pi\right)}{p\left(\mathbf{h}_\tau | \mathbf{h}_{\tau-1}, \pi\right)},
\end{aligned}
\tag{2.47}
$$

where the third term becomes a consistency term, as it can be read as the KL divergence between the forward message and the belief about a state conditioned on the respective observation. Under the Bethe approximation (Equation (2.29)) the predicted free energy can be decomposed as

$$
\begin{aligned}
G_\pi\left(\tau\right) = & \sum_{\mathbf{o}_\tau} q\left(\mathbf{o}_\tau, \mathbf{h}_\tau | \pi\right) \left[\ln \frac{q\left(\mathbf{o}_\tau | \pi\right)}{\bar{p}\left(\mathbf{o}_\tau\right)} - \ln p\left(\mathbf{o}_\tau | \mathbf{h}_\tau\right)\right] \\
& + \sum_{\mathbf{h}_\tau, \mathbf{h}_{\tau-1}} q\left(\mathbf{o}_\tau, \mathbf{h}_\tau, \mathbf{h}_{\tau-1} | \pi\right) \left[\ln \frac{q\left(\mathbf{h}_\tau, \mathbf{h}_{\tau-1} | \pi\right)}{q\left(\mathbf{h}_\tau | \pi\right) q\left(\mathbf{h}_{\tau-1} | \pi\right)} + \ln \frac{q\left(\mathbf{h}_\tau | \mathbf{o}_\tau, \pi\right)}{p\left(\mathbf{h}_\tau | \mathbf{h}_{\tau-1}, \pi\right)}\right],
\end{aligned}
\tag{2.48}
$$

where we recover an additional term compared to the mean-field approximation. This term corresponds to the mutual information between successive states, which defines the complexity cost of representing statistical dependence between hidden states. The final term in Equation (2.48) is the same as in the mean-field approximation, but it cannot be interpreted as easily here, as the messages under the Bethe approximation have a different form.

However, under the predicted free energy all terms but the norms of the messages cancel out (see Equation (2.36)) once the results for the approximate posterior are inserted. These norms can be interpreted as trial-dependent surprise, encoding the discrepancy between the forward planning and the prior expectations over future outcomes. With the predicted free energy, independent of the decomposition, the probability of reaching the goal state is the driving factor for agent behavior.

Importantly, simulating agent behavior using the expected rather than the predicted free energy leads to a relative tendency to choose paths towards states with low observation

uncertainty. When visiting these states, an observation is more informative about its underlying hidden state. An agent thereby reduces its uncertainty about its current state. In future work we will investigate if we can recover this information seeking behavior with the formalism based on the predicted free energy.

3 Balancing control: A Bayesian interpretation of habitual and goal-directed behavior

3.1 Abstract

In everyday life, our behavior varies on a continuum from either automatic and habitual to deliberate and goal-directed. Recent evidence suggests that habit formation and relearning of habits operate in a context-dependent manner: Habit formation is promoted when actions are performed in a specific context, while breaking off habits is facilitated after a context change. It is an open question how one can computationally model the brain's balancing between context-specific habits and goal-directed actions. Here, we propose a hierarchical Bayesian approach for control of a partially observable Markov decision process that enables conjoint learning of habit and reward structure in a context-specific manner. In this model, habit learning corresponds to a value-free updating of priors over policies and interacts with the value-based learning of the reward structure. Importantly, the model is solely built on probabilistic inference, which effectively provides a simple explanation how the brain may balance contributions of habitual and goal-directed control. We illustrated the resulting behavior using agent-based simulated experiments, where we replicated several findings of devaluation and extinction experiments. In addition, we show how a single parameter, the so-called habitual tendency, can explain individual differences in habit learning and the balancing between habitual and goal-directed control. Finally, we discuss the relevance of the proposed model for understanding specific phenomena in substance use disorder and the potential computational role of activity in dorsolateral and dorsomedial striatum and infralimbic cortex, as reported in animal experiments.

3.2 Introduction

In both psychology and neuroscience, theories postulate that behavioral control can vary along a dimension with habitual, automatic actions on one end, and goal-directed, controlled actions on the other (Wood & Rünger, 2016). In the context of operant conditioning, habits

have been described as retrospective and have been found to implement an automatic tendency to repeat actions which have been rewarded in the past (Dickinson et al., 1983; Graybiel, 2008). Habitual action selection is typically fast but is insensitive to outcomes and only slowly adapts to a changing environment (Seger & Spiering, 2011). In contrast, goal-directed action selection is prospective and implements planning based on a representation of action-outcome contingencies (Dickinson & Balleine, 1994; Dolan & Dayan, 2013). Consequently, goal-directed action selection adapts rather rapidly to a changing environment, but under a penalty of costly and slow computations.

Habit learning can be viewed as a transition from goal-directed to habitual behavior while a subject learns about its environment (Graybiel, 2008): In a novel environment or context, goal-directed actions will first allow the organism to learn about its structure and rewards and, later, to integrate this information to reliably reach a goal. With time, certain behaviors will be reinforced, while others will not. Subsequently, habits are formed to enable faster and computationally less costly selection of behavior which have been successful in the past. Given enough training, behavior is thought to be dominated by stimulus-driven habits, see e.g. (Dickinson, 1985; Seger & Spiering, 2011) for experimentally derived criteria of habit learning. In particular, two influential criteria are the insensitivity to contingency degradation where action-outcome associations are changed, and the insensitivity to reinforcer devaluation, where the outcome is made undesirable (Yin & Knowlton, 2006). Here, an established habit seems to make it difficult for an organism to change the previously reinforced habitual choice and adapt behavior to the altered conditions in its environment. Additionally, the strength of the habit and resulting insensitivity to changes has been found to critically depend on the duration and reward schedule of the training phase (Yin & Knowlton, 2006).

Importantly, habit learning as well as changing existing habits is strongly associated with the consistency of the environment while actions are performed (Wood & Rünger, 2016). When a specific behavior is executed in a stable context, habits are learned faster, and adjustment of behavioral patterns after changes in context is impeded (Lally et al., 2010). Conversely, learning of habits is slower and adjustment to changes is facilitated in a changing environment or inconsistent contexts. For example, it has been shown that learning of habits is improved when actions are mostly performed in the same context, e.g. after breakfast (Lally et al., 2010; Danner et al., 2008; D. T. Neal et al., 2012); while the unlearning of habits is improved after a context change, e.g. after a move to a different city (Verplanken & Roy, 2016).

In addition, habit learning trajectories strongly vary between individuals (Dolan & Dayan, 2013; Lally et al., 2010). Recent substance use disorder (SUD) studies show differences, between patients and controls, in learning and in the reliance on the so-called habit system, which lead to individual habitual responding biases (Ersche et al., 2016; Lim et al., 2019; Heinz et al., 2019). Still, it is an open question whether these different habit learning trajectories in individuals with SUD are due to individual factors or caused by the substance use itself (Nebe et al., 2018).

While there are findings that there are two hypothesized systems, the habitual and goal-directed system, and how they map onto brain structures (Dolan & Dayan, 2013; Yin & Knowlton, 2006; Everitt & Robbins, 2005), it is not clear if such a dichotomy is required for the computational description of these processes and for a mechanistic understanding of how habitual and goal-directed control are balanced, e.g. (Goschke, 2014). It has been argued that goal-directed and habitual behavior can be equated to model-based and model-free reinforcement learning (Dolan & Dayan, 2013). However, experimental evidence indicates that model-free reinforcement learning does not capture all experimentally established properties of habit formation (Friedel et al., 2014; Gillan et al., 2015). Rather, an alternative proposal is

centered on the idea that habits, as stimulus-response associations, may arise from repetition alone and are learned via a value-free mechanism (Miller et al., 2019). Another emerging research direction, built on both experimental and computational studies, is to consider habits as chunked action sequences, which may be modelled in a hierarchical fashion (Smith & Graybiel, 2016; Graybiel & Grafton, 2015; Dezfouli & Balleine, 2012, 2013; Graybiel & Grafton, 2015).

Here, we propose a hierarchical Bayesian habit-learning model based on the concept of planning as inference (Attias, 2003a; Botvinick & Toussaint, 2012a), which we will treat with methods of approximate inference (K. Friston et al., 2015). Critically, we regard habits as a prior over policies (sequences of actions), see also (K. Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty, & Pezzulo, 2016), which enables a novel way to understand how the brain may balance its action control between habitual and goal-direction contributions. In this model, the prior over actions is learned according to a Bayesian value-free update rule based on a tendency to repeat past actions. At the same time, the reward structure of the environment is learned in a value-based and outcome-sensitive manner. This learned reward structure is used for goal-directed action evaluation based on explicit forward planning which is computed in a likelihood. Action selection is implemented as sampling from the posterior which is the product of the prior and the likelihood, yielding an automatic balancing between goal-directed and habitual behavior. Importantly, habits and outcome rules are learned in a context-specific manner, and can be retrieved when revisiting a context. We use this hierarchical model to explain the transition dynamics from goal-directed to habitual behavior when learning habits, and adaptation of behavior to context changes.

In concrete terms, we propose to view balancing of behavioral control in a Bayesian way: Behavior is sampled from a posterior which, according to Bayes' rule, is a prior times a likelihood. We interpret the prior as the habit, where the habitual contribution for a specific action is higher the more this action, or sequence of actions, has been selected in the past. The goal-directed value of an action is encoded in the likelihood, where explicit forward planning yields the expected reward of an action. This explicit forward planning is based on learning of outcome contingencies, which allow the agent to predict the goal-directed value. As a result, the interpretation of how control is balanced is rather simple: Goal-directed and habitual value are multiplied using Bayes' rule, yielding an natural weighting of of their contributions to control based on the respective certainties. Importantly, the habit, i.e. the prior, and the outcome rules, and in effect the likelihood, are learned in a context-specific manner. As a result, habits and outcome contingencies are learned for each context and can be retrieved when re-encountering a known context.

We show that the proposed model is in principle able to capture basic properties of classical habit learning experiments: Insensitivity to changes in action-outcome contingencies and reinforcer devaluation, and the increase of this effect with longer training duration. We introduce a free parameter of the model, the habitual tendency, which modulates an individual's habit learning speed. We also show that stochastic environments which are akin to interval reward schedules result in an over-reliance on habitual control. Furthermore, we illustrate that context-specific habits enable rapid adaptation after a switch to another but already known context.

We will discuss the implications of our model and how the proposal of habits modelled as prior over action sequences lets us reinterpret the assumed dichotomy of the habitual and goal-directed system. In particular, we will briefly discuss the potential relevance of the impact of misguided context inference on the arbitration between habitual and goal-directed control in SUD and speculate on the mapping between specific model mechanisms and

recent findings in both the dorsolateral and dorsomedial striatum and the infralimbic cortex.

3.3 Methods

3.3.1 The generative process

In this work, we propose a hierarchical Bayesian model which implements context-dependent habit learning. We will describe the proposed modelling approach in detail and in a didactic fashion. Before we show details of the model, we describe the structure of the task environment. Our description rests on a hierarchical partially observable Markov decision process (POMDP), which is defined by the tuple $(\mathcal{S}, \mathcal{R}, \mathcal{A}, \mathcal{C}, \mathcal{T}_s, \mathcal{T}_r, \mathcal{T}_c)$, where

- $\mathcal{S} = \{s_1, \ldots, s_{n_s}\}$ is a set of states
- $\mathcal{R} = \{r_1, \ldots, r_{n_r}\}$ is a set of rewards
- $\mathcal{A} = \{a_1, \ldots, a_{n_a}\}$ is a set of actions
- $\mathcal{C} = \{c_1, \ldots, c_{n_c}\}$ is a set of contexts
- $\mathcal{T}_s \left(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t\right)$ is a set of action-dependent state transition rules
- $\mathcal{T}_r \left(\mathbf{r}_t | \mathbf{s}_t, \mathbf{c}_k\right)$ is a set of context-dependent reward generation rules
- $\mathcal{T}_c \left(\mathbf{c}_{k+1} | \mathbf{c}_k\right)$ is a set of context transition rules.

For a tutorial on POMDPs see (Littman, 2009). We partition the time evolution of the environment into N_e episodes of length T (Hommel, Müsseler, Aschersleben, & Prinz, 2001; Zacks, Speer, Swallow, Braver, & Reynolds, 2007; Butz, 2016). In the k-th episode, the environment is in context $\mathbf{c}_k \in \mathcal{C}$. In this episode, the first time step is $t = 1$. The environment starts out in its starting state $\mathbf{s}_1 \in \mathcal{S}$. Depending on the state and the current context, the environment distributes a reward $\mathbf{r}_1 \in \mathcal{R}$ according to the generation rule $\mathcal{T}_r$, which essentially encodes the contingency tables for each context. Note that a no-reward is also part of the set of rewards $\mathcal{R}$. This way, the environment is set up to have a context-dependent reward distribution rule, which may also change, when the environment transitions to a new context. Using these transitions, we will be able to implement the training and extinction phases of a typical habit learning environment as latent contexts in the Markov decision process.

A participant or agent, which is interacting with this environment, observes the reward and state of the environment, and chooses an action $\mathbf{a}_1$. This marks the end of the first time step $t = 1$ of the k-th episode. This process for a single time step is also shown in the left part of Figure 3.1.

In the second time step $t = 2$ of the k-th episode, the environment updates its state to a new state $\mathbf{s}_2$, in accordance with the context transition rule $\mathcal{T}_s$, depending on the previous state $\mathbf{s}_1$ and the chosen action $\mathbf{a}_1$. Given the new state and the current context, a new reward $\mathbf{r}_2$ is distributed. The agent once again perceives the state and reward and chooses a new action $\mathbf{a}_2$.

This process is iterated until the last time step $t = T$ of the episode is reached. In between the last time step of the current episode k, and the first time step of the next episode $k+1$, the context is updated to a new context $\mathbf{c}_{k+1}$ in accordance with the transition rule $\mathcal{T}_c$. Importantly, the context is an abstract, hidden (latent) state, which determines the current outcome rules

of the environment. It cannot be directly observed by the agent but only inferred from interactions with the environment. We chose this setup because in animal experiments the switch to the context of an extinction phase is typically not cued. Our assumption here is that an agent represents different environments with different rules as different contexts. As in daily life, rule changes might not be directly cued which makes it necessary to model uncertainty about context. This is in line with recent experiments and modelling work which demonstrated that humans and animals implicitly learn different outcome contingencies as different contexts, even when they are not cued (Palminteri, Khamassi, Joffily, & Coricelli, 2015; Gershman et al., 2010; Wilson, Takahashi, Schoenbaum, & Niv, 2014).

Note that this implementation effectively constitutes a hierarchical model on two different time scales: The episodes on the lower level, where states evolve quickly, i.e. in every time step, and the contexts on the higher level, which evolve more slowly, only every T time steps.

3.3.2 The generative model

To a participant or an artificial agent, this generative process is not directly accessible. Instead, the agent has to maintain a representation of this process, which is called the generative model. For the purpose of our model, we will assume that the agent knows which quantities are involved: It knows that there are states and that the possible states it could be in are summarized in the set $\mathcal{S}$. It also knows all possible rewards in $\mathcal{R}$, and all possible contexts in $\mathcal{C}$.

Furthermore, we assume that the principled structure of the environment is known to the agent: It knows that (i) state transitions depend on the previous state and the action chosen, (ii) reward generation depends on the current state and context, and (iii) the environment is partitioned into episodes, where the context is stable within but may switch between episodes. These causal relationships in the generative model are shown in Figure 3.2. Within an episode, we assume without loss of generality that the agent does not represent single actions, but sequences of actions (policies)

$$\pi = (\mathbf{a}_1, \ldots, \mathbf{a}_{T-1}) \in \{\pi_1, \ldots, \pi_{n_\pi}\}. \tag{3.1}$$

where a policy consists of $len(\pi) = T - 1$ actions because actions are executed in between time steps and an action at time step T would therefore have no effect.

Additionally we assume that the agent has the correct representation of the state transition rules $\mathcal{T}_s$. In other words, the agent knows which consequences its own actions will have. In contrast, we assume that an agent does not know the reward probabilities associated with each state and how they depend on the context. Instead, the agent represents those probabilities as random variables

$$\phi = \left\{\phi_{1,1,1}, \ldots, \phi_{r,s,c}, \ldots, \phi_{n_r,n_s,n_c}\right\} \tag{3.2}$$

which will have to be inferred.

Importantly, we propose that the agent learns context-dependent habits as a context-dependent prior over policies. It represents the parameters of this prior as latent random variables as well

$$\theta = \left\{\theta_{1,1}, \ldots, \theta_{\pi,c}, \ldots, \theta_{n_\pi,n_c}\right\}. \tag{3.3}$$

Formally, we write the causal structure of the agent's generative model as

$$p(\mathbf{s}_{1:T}, \mathbf{r}_{1:T}, \pi, \theta, \phi, \mathbf{c}_k) = p(\pi|\theta, \mathbf{c}_k)\, p\left(\theta|\alpha^{k-1}\right) p\left(\phi|\beta^{k-1}\right) p'(\mathbf{c}_k)\, p\left(\mathbf{s}_{1:T}, \mathbf{r}_{1:T}|\pi, \phi, \mathbf{c}_k\right) \tag{3.4}$$

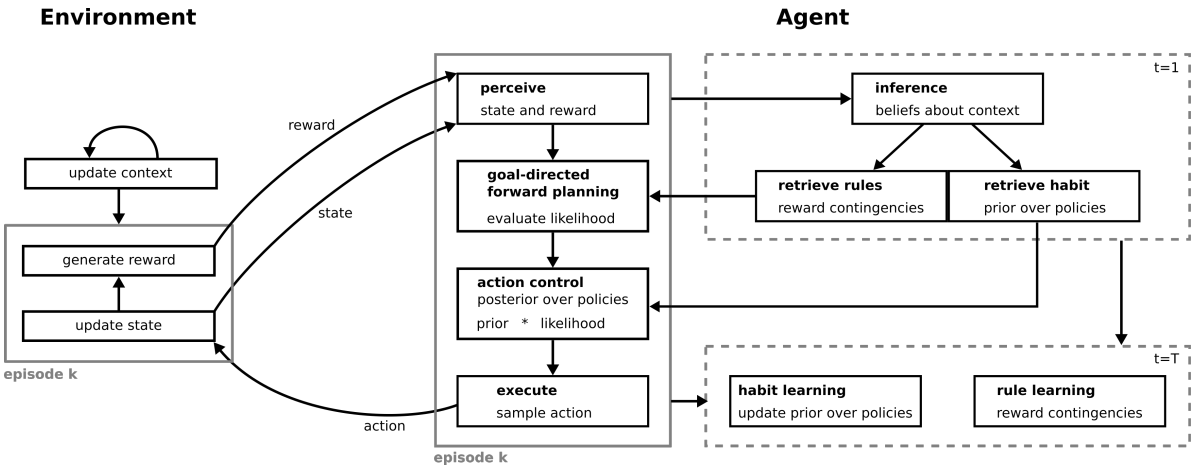

Figure 3.1: The agent in interaction with its environment

The environment (left) is modeled as a hierarchical partially observable Markov decision process (see Section 3.3.1). On the lower level, the time evolution of the environment is structured into episodes of length T. Here, the states of the environment evolve dependent on the previous state and action chosen by the agent. Given the state and the reward generation rules, some reward or no reward is distributed in each time step t of an episode. On the higher level, there is a slowly evolving context which determines the current rules of the environment, namely the reward generation rules, i.e. outcome contingencies. The agent (right) uses its generative model (see Section 3.3.2 and Figure 3.2) to represent the dynamics of the environment, and to plan ahead and select actions. At the beginning of each episode ($t = 1$), the agent infers the current context (box in the top right) based on previous rewards and states, and retrieves the learned reward generation rules and the habits (prior over policies) for this context. In each time step t in an episode, the agent perceives a new state-reward pair and uses forward planning in a goal-directed fashion (the likelihood) to then form a posterior over actions by combining the habit with the goal-directed computation what actions should be chosen. To execute an action, the agent samples from this posterior. This process repeats until the last time step $t = T$, where the agent updates its habits based on the policy it chose for this episode, and updates its knowledge about the reward structure based on the state-reward pairs it perceived (bottom right box). This updating is done in a context-specific manner so that the habits and rules are updated proportionally to the inferred probability of having been in a context during the past episode.

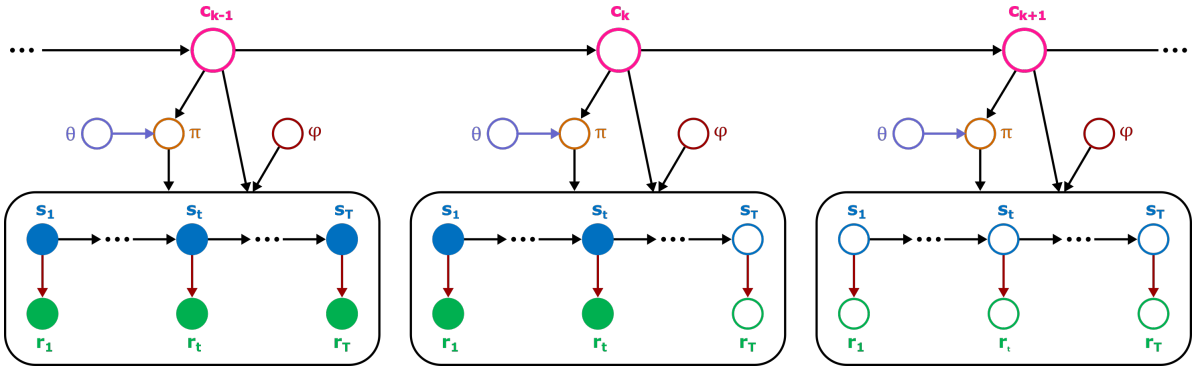

Figure 3.2: A graphical model depicting conditional dependencies between variables in the generative model

Empty circles indicate latent, unobservable variables and filled circles indicate known, observed variables, and arrows indicate statistical dependencies, where colored arrows indicate that these dependencies are learned by the agent. The model here is a hierarchical model, with the contexts $\mathbf{c}_k$ on the higher level of the hierarchy, and the episodes (black boxes) on the lower level of the hierarchy. In the current episode k (middle box), the agent starts at in some state $\mathbf{s}_1$ (blue), and receives a reward $\mathbf{r}_1$ (green) according to the current outcome rules (red downward arrows). The agent's knowledge about the current rules is represented by the parameters ϕ (red). The agent then chose some action $\mathbf{a}_1$ in accordance with a policy π (brown). For the next time step $t = 2$, the agent transitions to a new state $\mathbf{s}_2$ (arrow to the right), dependent on the policy π it followed (downward arrow from π), and a new reward $\mathbf{r}_2$ is distributed. This process repeated until the agent reached the current time step t. Viewed from here, all future states and rewards are unknown and, so far, unobserved variables, which the agent will infer during its planning process and evaluate if they lead to desirable outcomes. Based on this evaluation of the policies π and the prior over policies parameterized by θ (lilac), the agent can now choose a new action $\mathbf{a}_t$. On the higher level of the hierarchy, there are the latent contexts $\mathbf{c}_k$ (pink), which evolve more slowly (arrows to the right). They also determine which outcome rules are currently in use (downward right tilted arrow), and which prior over policies is being learned (downward left tilted arrow). The prior over policies is parameterized with the parameters θ (lilac), whose influence on the policy is also subjected to learning (lilac arrow to the right). We furthermore show the previous context $\mathbf{c}_{k-1}$ and the next context $\mathbf{c}_{k+1}$, which encode the previous episode (left box) and the next episode (right box), respectively.

where

$$p\left(\mathbf{s}_{1:T}, \mathbf{r}_{1:T} | \phi, \mathbf{c}_k\right) = \prod_{m}^{t} p\left(\mathbf{s}_m | \mathbf{s}_{m-1}, \pi\right) p\left(\mathbf{r}_m | \mathbf{s}_m, \phi, \mathbf{c}_k\right) \prod_{\tau=t+1}^{T} p\left(\mathbf{s}_\tau | \mathbf{s}_{\tau-1}, \pi\right) p\left(\mathbf{r}_\tau | \mathbf{s}_\tau, \phi, \mathbf{c}_k\right) p\left(R = 1 | \mathbf{r}_\tau\right)$$

is the agent's representation of the k-th episode, in which it is at time step t. This is an effective partition of states and rewards into past observed states $\mathbf{s}_{1:t}$ and rewards $\mathbf{r}_{1:t}$ and unknown future states $\mathbf{s}_{t+1:T}$ and rewards $\mathbf{r}_{t+1:T}$. The past states and rewards have been observed and are therefore known exactly to the agent. Conversely, the future states and rewards are unknown and are therefore latent variables which will have to be inferred. Note that this is an exact representation of the graphical model in Figure 3.2.

We use the following distributions to define the generative model:

- The policies π are represented by a categorical distribution

  $$p\left(\pi = l | \theta, \mathbf{c}_k = n\right) = \prod_{n,l} \theta_{l,n}^{\delta_{l,\pi}\delta_{n,\mathbf{c}_k}}$$

  where $\delta_{i,j}$ is the Dirac delta.

- The latent parameters of the prior over policies θ are distributed according to the respective conjugate prior, a product of Dirichlet distributions

  $$p\left(\theta | \alpha\right) = \prod_{n} Dir(\alpha_n^{k-1}) = \prod_{n} \frac{1}{B\left(\alpha_n^{k-1}\right)} \prod_{l} \theta_{l,n}^{\alpha_{l,n}^{k-1}-1}$$

  .

- The so-called concentration parameters $\alpha^{k-1} = \left\{\alpha_{l,n}^{k-1}\right\}$ are pseudo counts of the Dirichlet distributions. They encode how often an agent has chosen a policy in a specific context up until the previous episode $k-1$, and therewith shape the prior over policies.

- The rewards $\mathbf{r}_t$ are distributed according to a conditional categorical distribution

  $$p\left(\mathbf{r}_t = i | \mathbf{s}_t = j, \phi, \mathbf{c}_k = n\right) = \prod_{i,j,n} \phi_{i,j,n}^{\delta_{i,\mathbf{r}_t}, \delta_{j,\mathbf{s}_t}\delta_{n,\mathbf{c}_k}}$$

  .

- As above, the latent parameters ϕ are distributed according to the product of conjugate Dirichlet priors

  $$p\left(\phi | \beta\right) = \prod_{j,n} Dir(\beta_{j,n}^{k-1}) = \prod_{j,n} \frac{1}{B\left(\beta_{j,n}^{k-1}\right)} \prod_{i} \phi_{i,j,n}^{\beta_{i,j,n}^{k-1}-1}$$

  .

- The concentration parameters $\beta^{k-1} = \left\{\beta_{i,j,n}^{k-1}\right\}$ are pseudo counts of the Dirichlet distribution. They encode how often the agent saw a specific reward in a specific state and context up until the previous episode $k-1$. Therewith they represent the agent's knowledge about the reward generation rules, i.e. contingencies.

- The states are distributed according to a conditional categorical distribution

$$p\left(\mathbf{s}_t = j' | \mathbf{s}_{t-1} = j, \pi = l\right) = \prod_{j',j,l} p_{j',j,l}^{\delta_{j',\mathbf{s}_t},\delta_{j,\mathbf{s}_{t-1}},\delta_{l,\pi}}.$$

  We will fix the parameters $p_{j',j,l}$ to the true (deterministic) state transitions $\mathcal{T}_s$ in the generative process.

- The contexts are distributed according to a categorical distribution $p'\left(\mathbf{c}_k\right)$. We define this as a predictive prior $p'\left(\mathbf{c}_k\right) = p\left(\mathbf{c}_k | \mathbf{s}_{1:k-1}, \mathbf{r}_{1:k-1}\right)$ based on observed past states and rewards. Note that it also includes the agent's expectation of temporal stability of its environment. Specifically, we assume all contexts have the same temporal stability and change equally often.

- The agent's preference of rewards is represented by $p\left(R = 1 | \mathbf{r}_\tau\right)$, using a dummy variable R, see (Solway & Botvinick, 2012). High values of the probability distribution mean high preference for a particular reward, while low values mean low preference.

After having set up the generative model, we will now show how the agent, based on this model, forms beliefs about its environment and selects actions. To describe action evaluation and selection, we will follow the concept of planning as inference (Attias, 2003a; Botvinick & Toussaint, 2012a) and active inference (K. Friston et al., 2015; K. Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty, & Pezzulo, 2016; Schwöbel et al., 2018). Critically, this means that, apart from forming beliefs about hidden variables of the environment, actions or policies are also treated as latent variables that can be inferred.

3.3.3 Approximate posterior

When an agent infers hidden variables of its environment, such as the context, or future states and rewards, it needs to calculate the posterior

$$p\left(\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T}, \pi, \theta, \phi, \mathbf{c}_k | \mathbf{s}_{1:t}, \mathbf{r}_{1:t}\right) \tag{3.5}$$

over these hidden variables using Bayesian inversion. Intuitively this means asking the questions: What context am I most likely in, given I was in these states and received those rewards? What states will I visit in the future, and what rewards will I receive, given I have been in these states in the past and received those rewards? What are the most likely outcome rules that have generated rewards from states? To ensure analytical tractability and low computational costs, we will use variational inference as an approximate Bayesian treatment of the inference process.

Variational inference makes the inference process analytically tractable by replacing the computation of the true posterior with a simpler approximate posterior. In our case we will express the approximate posterior as

$$\begin{aligned} p\left(\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T}, \pi, \theta, \phi, \mathbf{c}_k | \mathbf{s}_{1:t}, \mathbf{r}_{1:t}\right) \approx & q\left(\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T}, \pi, \theta, \phi, \mathbf{c}_k\right) \\ = & q\left(\pi | \mathbf{c}_k\right) q\left(\theta | \alpha^k\right) q\left(\phi | \beta^k\right) q\left(\mathbf{c}_k\right) q\left(\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T} | \pi, \mathbf{c}_k\right) \end{aligned}$$

where we use belief propagation based on the Bethe approximation within a behavioral episode

$$q\left(\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T} | \pi, \mathbf{c}_k\right) = \prod_{\tau=t+1}^{T} \frac{q\left(\mathbf{s}_\tau, \mathbf{s}_{\tau-1} | \pi, \mathbf{c}_k\right)}{q\left(\mathbf{s}_\tau | \pi, \mathbf{c}_k\right)} \frac{q\left(\mathbf{r}_\tau, \mathbf{s}_\tau | \pi, \mathbf{c}_k\right)}{q\left(\mathbf{s}_\tau | \pi, \mathbf{c}_k\right)} \tag{3.6}$$

This is well motivated because within an episode, states and rewards critically depend on each other so it is sensible to use an approximation which captures these dependencies.

Outside of an episode, statistical dependencies may be averaged out, so that a mean-field approximation is sufficient to approximate the posterior. Specifically, we will use forward mean-field belief propagation, to obtain an agents beliefs based on the observed states and rewards. The posteriors of all random variables will be distributed the same way as in the generative model: states, rewards, policies, and context follow a categorical distribution; while their parameters θ and ϕ follow a Dirichlet distribution. These come out naturally from calculating the update equations (see Appendix).

3.3.4 Update equations

The marginal and pairwise approximate posteriors can be analytically calculated at the minimum of the variational free energy, see e.g. (Bishop, 2006a; Yedidia, Freeman, & Weiss, 2003b). These posteriors are typically called beliefs, as they encode the agent's beliefs about the hidden variables in its environment. We will now show the update equations resulting from the free energy minimization. These equations implement the agent's information processing: how it forms beliefs about the hidden variables in its environment, how it learns, plans, and evaluates actions. An illustration of this process is shown on the right side of Figure 3.1.

At the beginning of time step t in the k-th episode, the agent perceives the state $\mathbf{s}_t$ of its environment, and receives a reward $\mathbf{r}_t$. It uses this co-occurrence of state and reward to infer the current context and to update its beliefs about the reward generation rules. The posterior over context is estimated as

$$q\left(\mathbf{c}_k\right) = p'\left(\mathbf{c}_k\right) \exp\left(-F\left(\mathbf{c}_k\right)\right); \qquad p'\left(\mathbf{c}_k\right) = \sum_{\mathbf{c}_{k-1}} p(\mathbf{c}_k|\mathbf{c}_{k-1}) q(\mathbf{c}_{k-1}) \tag{3.7}$$

where $p'\left(\mathbf{c}_k\right)$ is a predictive probability for contexts given the beliefs previous episode and the transition probabilities $p(\mathbf{c}_k|\mathbf{c}_{k-1})$, and $F\left(c_k\right)$ is the context-specific free energy. The free energy term $F\left(c_k\right)$ encodes the approximate surprise of experienced rewards, states, and the agent's actions in different possible contexts (see Appendix). The more expected the rewards and actions are for a context, the lower this free energy, and the higher the posterior probability which the agent assigns to this context. As a result, an agent will infer to be in a stable context as long as rewards and actions are as expected, while it will infer a context change if outcomes and actions are unexpected. Note that, initially, before encountering any context, the prior over contexts $p'\left(\mathbf{c}_1\right)$ cannot be set to be uniform. It needs to have a bias towards one of the contexts, so that the agent knows to associate the experienced reward contingencies with the respective context. Which context is assumed to come first is not important, but we found that the agent's (intuitive) belief that it is most likely in some context is essential for the learning process.

The posterior beliefs about the reward probabilities are again a product of Dirichlet distributions, whose parameters are updated as

$$\begin{aligned} q\left(\phi|\beta^k\right) &= \prod_{j,n} \frac{1}{B\left(\beta^k_{j,n}\right)} \prod_i \phi_{ijn}^{\beta^k_{ijn}-1} \\ \beta^k_{ijn} &= \beta^{k-1}_{ijn} + q\left(\mathbf{c}_k = n\right) \sum_{m=1}^{t} \delta_{i,\mathbf{r}_m} \delta_{j,\mathbf{s}_m} \end{aligned} \tag{3.8}$$

which corresponds to updating pseudo counts β_{ijn}^k. The pseudo counts help keep track of how often the agent has seen a specific reward i in a specific state j and context n. Each time a new reward is generated in a state, these counts are increased by $q\left(\mathbf{c}_k\right)$. This way, the counts are high for context with high posterior probability and corresponding observed sequence of reward-state pairs, and low otherwise. At the beginning of a new episode, this posterior will become the new prior, which corresponds to a learning rule in between episodes.

The agent can now use its new knowledge about the rules of its environment to plan into the future and evaluate actions based on their expected outcomes. In order to plan ahead, it calculates its beliefs about future states $q\left(\mathbf{s}_\tau\right)$ and resulting future rewards $q\left(\mathbf{r}_\tau\right)$ in the current episode. These beliefs are calculated using belief propagation update rules (see Appendix). If a policy π predictably leads to states which yield desirable rewards, as encoded by the outcome preference $p\left(R=1|\mathbf{r}_\tau\right)$, this policy has a low policy-specific free energy (low surprise) $F\left(\pi|\mathbf{c}_k\right)$. The posterior beliefs over policies are computed as

$$q\left(\pi|\mathbf{c}_k\right) \propto p'\left(\pi|\mathbf{c}_k\right) \exp\left(-F\left(\pi|\mathbf{c}_k\right)\right); \qquad \ln p'\left(\pi|\mathbf{c}_k\right) = \int d\theta q\left(\theta\right) \ln p\left(\pi|\theta, \mathbf{c}_k\right) \tag{3.9}$$

where the free energy corresponds to the log-likelihood in a simple Bayes equation. Importantly, the log-likelihood represents the agent's goal-directed, value-based evaluation of actions, as it assigns them a value based on predicted future rewards. Additionally, the posterior beliefs contain a prior $p'\left(\pi|\mathbf{c}_k\right)$, which assigns an a priori weight to different policies or actions (Doshi-Velez, Wingate, Roy, & Tenenbaum, 2010; Todorov, 2009; K. Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty, & Pezzulo, 2016). In our work, this prior plays an important role, as we propose to interpret this prior as the habit of an agent. This is well motivated, because such a context-specific prior implements a planning-independent, i.e. value-free, tendency to choose an action (Miller et al., 2019). The agent then samples its next action from the posterior above, which is the product of the prior times the likelihood. Critically, this leads to an automatic weighting, i.e. arbitration, between goal-directed control (the likelihood) and habitual control (the prior) of the agent's next action.

At the end of an episode, after having sampled a policy and executed the respective actions, the agent updates its posterior beliefs about the prior over policies

$$\begin{aligned} q\left(\theta|\alpha^k\right) &= \prod_n \frac{1}{B\left(\alpha_n^k\right)} \prod_l \theta_{ln}^{\alpha_{ln}^k - 1} \\ \alpha_{ln}^k &= \alpha_{ln}^{k-1} + q\left(\pi = l|\mathbf{c}_k = n\right) q\left(\mathbf{c}_k = n\right) \end{aligned} \tag{3.10}$$

which constitutes habit learning in our model. Here, the pseudo counts α_{ln}^k are increased when a policy is chosen in a specific context. After the episode, this posterior becomes the new prior, in order to enable learning across episodes. Note that this implements a tendency to repeat previous actions on one hand, but also to repeat behavior which has been successful in the past. While the prior is independent from the goal-directed evaluation in the likelihood, it is based on which policies were previously chosen. This in turn is influenced by the goal-directed evaluation at the time when they were chosen. In other words, the habit and the outcome rules are learned conjointly. This is an important point because it means that goal-directed control and habit learning are intertwined in a specific way, see also Discussion.

The way the policy pseudo counts α_{ln}^0 are initialized before the first interaction with any context plays a critical role in how an agent learns a habit. Low initial counts $\alpha_{ln}^0 = \alpha_{\text{init}} = 1$ (for every l, n) mean that each time a new policy is chosen in a context, the pseudo count increases by a value between 0 and 1 (the posterior over contexts), which increased the count

substantially. As a result, the prior over policies becomes fairly pronounced very quickly. In contrast, a high initial count $\alpha_{\text{init}} = 100$ means that habits are learned a lot slower, as adding one to this value will have little influence on the prior probability of the corresponding policy. Therefore, we will define a habitual tendency as

$$h = \frac{1}{\alpha_{\text{init}}} \in [0, 1] \tag{3.11}$$

which we will consider a free model parameter with respect to which we will investigate behavioral differences. A high habitual tendency close to 1 will lead to an agent being a strong habit learner and exhibiting fast habit acquisition, while a low habitual tendency close to 0 will lead to a a weak habit learning with a low habit learning rate.

3.3.5 Simulation analyses

In this section, we will define quantities which we will use to illustrate our results. Specifically, we will want to investigate how agents infer contexts, using the posterior over contexts $q(\mathbf{c}_k)$, and how agents choose actions, using the marginalized posterior over policies

$$q(\pi) = \sum_{\mathbf{c}_k} q\left(\pi|\mathbf{c}_k\right) q(\mathbf{c}_k) \tag{3.12}$$

Specifically, to replicate standard results from experimental research, we will report simulations in an environment with two contexts $\mathcal{C} = \{c_1, c_2\}$ and two actions $\mathcal{A} = \{a_1, a_2\}$. We set episodes to length $T = 2$, so that actions and policies map one to one, which corresponds to a planning depth of 1. We use such short episodes here so that an episode is equivalent to one trial in a habit learning experiment. Nonetheless, it is possible to have longer episodes with increased planning depth in this model, which would endow an agent with the opportunity to learn habits as sequences of actions (see Discussion).

As we have binary random variables, for both contexts and actions we can completely capture the posterior beliefs with a single quantity, the posterior probability of being in second context ($Q_c := q(\mathbf{c}_k = c_2) \in [0, 1]$) and the posterior probability of selecting the second option ($Q_a := q(\pi = a_2) \in [0, 1]$). The posterior probability of being in first context, or selecting first option are obtained as $1 - Q_c$, and $1 - Q_a$, respectively.

In a similar vein, we also define the likelihood $L_a(k) := \sum_{\mathbf{c}_k} q(\mathbf{c}_k) \exp\left(-F\left(\pi = a_2|\mathbf{c}_k\right)\right)/Z_c$ of the second option in order to illustrate the agents goal-directed system, and the prior $P_a(k) := \sum_{\mathbf{c}_k} q(\mathbf{c}_k)\, p'\left(\pi = a_2|\mathbf{c}_k\right)$ to illustrate how an agent learns habits. The environment will be set to context 1, in a training phase, and switched to context 2 in an extinction phase. When the context switches, the posterior probabilities Q_c, and Q_a should transit from being close to zero, to being close to one, expressing changes in the posterior beliefs as a consequence of the changes in the underlying latent variables. Hence, we assume that the belief trajectory can be fitted with a sigmoid function

$$Q_a(k), Q_c(k) \approx \sigma\left(k|\gamma^{a,c}\right) = \frac{\gamma_1^{a,c}}{1 + \exp\left(-\gamma_2^{a,c}\left(k - \gamma_3^{a,c}\right)\right)} + \gamma_4^{a,c} \tag{3.13}$$

The motivation for this approximation of the trajectory is to determine the trial or episode (k^*) at which posterior beliefs Q_c, and Q_a transit from close to 0 to close to 1. The inflection point is specified by the parameters γ_3^c and γ_3^a, for Q_c and Q_a respectively. We have used the

implementation from Python3 SciPy $1.1.0$ (Virtanen et al., 2019) of nonlinear curve fitting for this procedure.

We also define a habit strength H to quantify the strength of habitual control under different conditions. We define the habit strength as the delay between the actual switch in context of the environment, and the time point at which an agent adapts their behavior. The change in context in our experiment relates to the switch between the training and extinction phases. The time point of adaptation can be interpreted as the trial in which the posterior over actions flips from close to 0 to close to 1. This equates to the inclination point of the sigmoid fitted to the posterior over actions. We define the habit strength as

$$H = \gamma_3^a - d_{\text{training}} \in [1, 100] \tag{3.14}$$

as the difference between the fitted inclination point γ_3^a and the training duration d_{training}. The extinction phase in which we will test for habitual behavior will have 100 trials. As a result, the habit strength can be between 1 and 100, where $H = 1$ indicates that an agent immediately switched its behavior in the first trial of the extinction phase and showed no habitual control, while $H = 100$ means that an agent failed to adapt within the extinction phase and therewith showed full habitual control.

We used the implementation of t-test and ANOVA provided by the Scipy $1.1.0$ (Virtanen et al., 2019) package. Similarly, we performed the linear regression the implementation of the ordinary least squares (the OLS class) provided in the StatsModels $0.10.1$ (Seabold & Perktold, 2010) package.

3.4 Results

Having derived the update equations of the proposed model, we will now use a series of simulated experiments to show how an artificial agent controls its behavior by balancing between habitual and goal-directed control. In these simulations, we will use environments where agents are required to adapt their behavior to context switches. In Section 3.4.1, we will first introduce a task which captures key features of habit learning similar to animal experiments, specifically contingency degradation and outcome devaluation, where we test for habitual behavior in extinction. We will present six different results:

- We let two exemplary agents perform the task under contingency degradation, show internal properties of the model, and how agents learn habitual behavior (Section 3.4.2).
- We demonstrate how internal model parameters, like the habitual tendency $h = \frac{1}{\alpha_{\text{init}}}$, influence the agent's information processing, behavior, and that an increased habitual tendency increases habit strength after contingency degradation (Section 3.4.3).
- We show that the acquired habit strength depends on training duration (Section 3.4.4).
- We show a specific advantage of contextual habit learning, namely that contextual habits allow optimized behavior to be retrieved quickly, when an agent is revisiting a previously experienced context (Section 3.4.5).
- We show how environmental stochasticity, e.g. highly probabilistic rewards, leads to an over-reliance on habitual behavior and increase habit strength (Section 3.4.6).
- We introduce outcome devaluation to the task and show that agents exhibit habitual behavior insensitive to contingency degradation and outcome devaluation (Section 3.4.7).

3.4.1 Habit learning task

A common way to experimentally test for habit formation in animal experiments is contingency degradation (Yin & Knowlton, 2006; Wood & Rünger, 2016). Here, an animal is probabilistically rewarded after performing a specific action, e.g. pressing a lever. After a training period, in which the animal learns action-outcome associations and potentially acquires a habit, habitual behavior is measured in an extinction period. The outcome contingencies of the environment are changed, and the lever press does not yield a reward any longer. Conversely, the animal is often rewarded for abstaining from pressing the lever. After this change of contingencies, the strength of habitual control is assessed as the continuation of lever pressing, where a higher habit strength corresponds to more presses. For moderate training durations ($\sim 50 - 100$ trials), the animal will have formed a weak or no habit, and seizes to press the lever rather quickly. For extensive training (~ 500 trials), experiments show that the animal will have formed a strong habit and will continue to press the lever for an extended period of time (~ 50 trials), e.g. (Colwill & Rescorla, 1988; Adams, 1982).

Additionally, for behavior to be classified experimentally as habitual, it must be insensitive to outcome devaluation (Yin & Knowlton, 2006). Here, animals undergo a similar training as in contingency degradation experiments. Then, outcomes are devalued by either satiating the animals, or by associating the reinforcer with an aversive outcome. Afterwards, behavior is again tested in extinction, where a continuing of the lever press is interpreted as evidence for habitual behavior, see e.g. (Adams, 1982). Typically, the strength of habitual behavior also greatly depends on the reinforcement schedule (Yin & Knowlton, 2006), which may be a ratio schedule, where each action leads to a reward with a specific probability, or an interval schedule, where rewards are only distributed after a certain time has elapsed. Interval schedules lead to a greater habit strength and decreased sensitivity to changes in outcome contingencies.

To demonstrate that the proposed model can replicate these basic features of habit learning, we approximate the experimental setup of a habit learning experiment in a simplified way, by using a so-called two-armed bandit task, see Figure 3.3a. This way of modelling the task follows previous modelling studies such as (Daw et al., 2005; S. W. Lee, Shimojo, & O'Doherty, 2014) and emulates probabilistically rewarded lever presses of the animal. In the proposed habit learning task, an artificial agent can choose to perform either action a_1, i.e. press a left lever 1, or action a_2 to press a right lever 2. Each lever pays out a reward according to the reward generation rules $\mathcal{T}_r$, and these probabilities will switch after certain number of trials, emulating a contingency change, similar to habit learning experiments (Figure 3.3b). In many habit learning experiments, the animals do not choose between two levers, but rather between pressing a lever or abstaining from pressing, where abstaining is a viable option due to opportunity costs. We approximated opportunity costs of not pressing the lever by introducing a minimally rewarded second choice (lever 2) instead, see also similar approaches taken in previous modelling studies (Daw et al., 2005; S. W. Lee et al., 2014; Keramati, Dezfouli, & Piray, 2011; Pezzulo, Rigoli, & Chersi, 2013; Gershman, Markman, & Otto, 2014).

The habit learning task has two phases (Figure 3.3b): The first phase is the training phase which lasts $d_{\text{training}} = 100$ trials. We will also vary this duration in Section 3.4.4. Here, lever 1 pays out a reward with $\nu = 0.9$ probability, and lever 2 with $1 - \nu = 0.1$. These reward probabilities are kept stable during the training period and the agent learns about outcome contingencies and might form a habit. The second phase is the extinction phase which lasts another 100 trials. Here, outcome probabilities are switched relative to the training phase, and are kept stable for the remainder of the experiment. After the switch of outcome

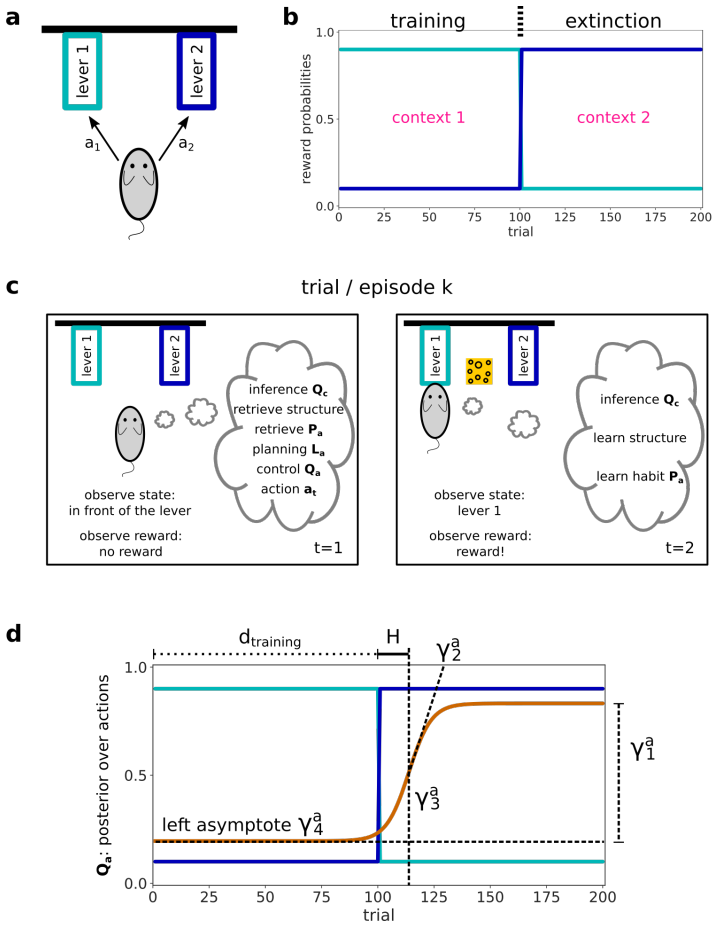

Figure 3.3: Habit learning task

a In each trial k, the agent can choose between pressing two levers (light and dark blue boxes, lever in black next to the box) and is awarded probabilistically. We model this task as a two-armed bandit task. **b** Reward schedule over 200 trials for the two levers. In the training phase, lever 1 yields a reward with $\nu = 0.9$ probability, while lever 2 only yields a reward with $1 - \nu = 0.1$ probability. After 100 trials, the reward probabilities switch. The new contingencies are stable for another 100 trials. This second stable period emulates an extinction phase, where we will test the agent's habit strength by how quickly it is able to adapt its choices. **c** An agent solving the task. For the agent, each trial constitutes one behavioral episode. In episode or trial k, the agent starts out in the state (position) in front of the two levers in the first time step $t = 1$ of this episode. It observes its state and that there is no reward. The agent can now infer the context Q_c based on its experience in the previous trials. It retrieves the learned outcome contingencies and habit P_a for this context from memory. It uses its knowledge about the reward structure to plan forward and evaluate actions based on the likelihood L_a, where actions which lead more likely to a reward will have a higher likelihood encoding the goal-directed value. The agent combines the likelihood and the prior to evaluate the posterior over actions Q_a and samples a new action $\mathbf{a}_t$ from this posterior, for example action a_1. In between episodes, this action is executed and the agent transitions to the new state, pressing lever 1. At the beginning of the next time step $t = 2$, a reward may be distributed, depending on the action and lever the agent chose. It then updates its context inference Q_c based on the perceived state-reward pair, learns the outcome rules, and updates its habit P_a. This process repeats until the last trial $k = 200$. **d** Illustration of the sigmoid function used to analyse the time evolution of the posterior over actions Q_a (see Section 3.3.5 for details). The a as a superscript on the parameters signifies that these are the parameters for the posterior over actions. We define the habit strength H as the difference between the inflection point of the posterior beliefs (γ_3^a) and the trial number at which the context changed d_{training}.

contingencies, we quantify an agent's habit strength as the number of trials before an agent adapts its behavior and primarily presses lever 2 instead of lever 1, see section 'Simulation analyses' in Methods. Note that in our simulations, due to our agent setup, a trial is equivalent to a behavioral episode for an agent, see Figure 3.3c for an exemplary episode in which the agent interacts with the habit learning task.

This experimental setup emulates the training and extinction phases of a contingency degradation habit learning experiment. It can be transformed into a outcome devaluation experiment by modulating the agent's preference for outcomes ($p\left(R=1|\mathbf{r}_t\right)$, see Section 3.3.2 and Appendix) after the training phase. In order to disentangle these two effects, we will restrict our simulated experiments to contingency degradation in most of the following sections. In the last section, we will show habitual behavior under outcome devaluation.

Note that the two phases of the experiment (Figure 3.3b) can be viewed as a sequence of two contexts, where in each context one of the two choices returns higher expected reward. Importantly, the agent is initially not explicitly aware how any context is associated with a specific set of outcome rules. Instead, the agent learns to associate the outcome rules it first experiences with the first context. When the contingencies change, it will infer the change and learn to associate the new rules with a second context. By design in our experiment, this corresponds to associating contexts with preferable levers. In some habit learning experiments, contexts are cued and habitual behavior is used in response as form of stimulus response association, e.g. (Sage & Knowlton, 2000). In our habit learning task, we do not use a cue to indicate the context to the agent. This is in line with typical animal experiments where the extinction phase is not cued. Instead, the state, i.e. the position of the agent in front of the levers is observable and takes the role of a stimulus.

3.4.2 Habit learning under contingency degradation

In this section, we illustrate, in detail, how agents based on the proposed model learn about their environment, form beliefs, acquire habits, select actions, and balance goal-directed and habitual control, see Methods and Figure 3.1. As the habitual tendency parameter h has a strong influence on habit learning and action selection, we will show two exemplary simulations of a an agent with strong ($h=1.0$) and another agent with a weak ($h=0.01$) habitual tendency performing the task (Figure 3.3). In the following, we refer to these two agents as the strong habit learner ($h=1.0$) and the weak habit learner ($h=0.01$). Note that, in this section, for didactic purposes, we will describe model behavior on just single instances of two representative agents. This is followed by more thorough simulations, where we also quantify the uncertainty over model variables using multiple experiments for each agent.

When an agent is first put into the task environment, it has no prior knowledge about the outcome contingencies associated with any context, and no prior preference for any actions $p'\left(\mathbf{a}_1|\mathbf{c}_1=c_1/c_2\right)=\left(\frac{1}{2},\frac{1}{2}\right)^T$, i.e. there is no habit yet. What the agent does know, is that action 1 means pressing lever 1, and action 2 means pressing lever 2, so that it has an accurate representation of the state transition matrices $p\left(\mathbf{s}_{t+1}|\mathbf{s}_t,\mathbf{a}_t\right)$. Furthermore, the agent has a prior over contexts with a bias towards context 1 (see Methods).

In the first trial, the agent has not sampled any reward yet, so it chooses an action $\mathbf{a}_1$ randomly as it does not have any knowledge available to predict the outcome of actions. According to the action chosen, the agent goes to and presses the respective lever, and receives a reward or no reward. At the end of the trial, as this also marks the end of a behavioral episode, the agent updates its prior P_a to increase the a priori probability to repeat

this chosen action, and updates its knowledge about the reward structure (see Figure 3.1 and Figure 3.3c). As the agent started with a biased prior over contexts, it associates this reward structure with context 1. Hence, the prior bias for context 1 simply reflects agent knowledge that it can be in only one context initially.

At the start of the second trial, the agent infers that it is most likely in context 1 (Q_c), based on its previous experience and its knowledge about the stability of the environment. It retrieves the reward structure and the prior P_a over actions it just learned. The agent can now use this new knowledge about outcome contingencies in the current context to evaluate the likelihood L_a. In order to select an action, it calculates the posterior beliefs over actions Q_a as the product of the prior P_a, which represents habits as an automatic and value-free tendency to repeat actions, and the likelihood L_a, which represents the goal-directed and value-based evaluation of anticipated future rewards (see Eq. 3.9). The agent then samples an action $\mathbf{a}_2$ from these posterior beliefs about actions, dynamically adjusting the balance between goal-directed and habitual choices. The agent visits and presses the lever it just chose and samples a reward. At the end of this trial and behavioral episode, the agent reevaluates its beliefs about the context Q_c, based on if the new observations still fit to its knowledge about this context. The agent also updates its prior over actions P_a (the representation of a habit), hence increasing the prior probability of that action being repeated. Similarly, the agent updates its knowledge about the reward structure, based on its beliefs about the context. This update cycle is repeated over all future trials, see Figure 3.3c and Section 3.3.4.

Figure 3.4 shows the resulting dynamics of the relevant agent variables (Q_c, L_a, P_a, Q_a, $\mathbf{a}_k$) for the strong (left) and weak (right column) habit learner during all 200 trials in the habit learning task. In the training phase, the beliefs over context Q_c converge rather quickly and after about 10 trials, the two agents are certain of being in context 1 (see Figure 3.4a). Figure 3.4b) shows the likelihood over actions L_a, reflecting the expected choice value, that is, the estimated surprise in reaching a goal (observing a rewarding outcome). As the likelihood depends on the learned knowledge about the environment, it takes both weak and strong habit learners around 30 trials to observe enough outcomes before the likelihood converges to a stable value. Figure 3.4c) shows the prior over actions P_a, i.e. the representation of a habit. Here, the difference between the strong and weak habit learner is obvious: The strong habit learner (left) forms a strong habit quickly ($P_a < 0.1$) after only 40 trials. This means, the strong habit learner has a very high a priori probability $1 - P_a$ of choosing action 1 independent of the expected rewards. Conversely, the weak habit learner updates its prior over actions rather slowly ($P_a \in [0.4, 0.6]$). The second to last row (Figure 3.4d)) shows the posterior over actions Q_a, which is the product of the prior and the likelihood. For the weak habit learner, the prior has little to no influence, as it is close to 0.5, so that the posterior over actions looks similar to the likelihood. For the strong habit learner, the strong prior lets the posterior over actions converge to values close to 1.0 within 40 trials. The agents sample their actions from this posterior probability, which are shown in the bottom row (Figure 3.4e)). The strong habit learner chooses the action with the higher expected reward more consistently (94% of choices), while the weak habit learner continues to choose action 2 even late into the training period. As a result, the weak habit learner has a significantly lower success rate (80%, $p = 0.003$, two sample t-test on the chosen actions in the training phase of two agents shown here).

In the extinction phase, after the switch in trial 100, the reward contingencies become reversed. When continuing to press lever 1, the agents are only rewarded with a probability of $1 - \nu = 0.1$. The lack of expected reward payout produces a prediction error which increases the context-specific free energy (see Section 3.3.4). This drives the agents to quickly

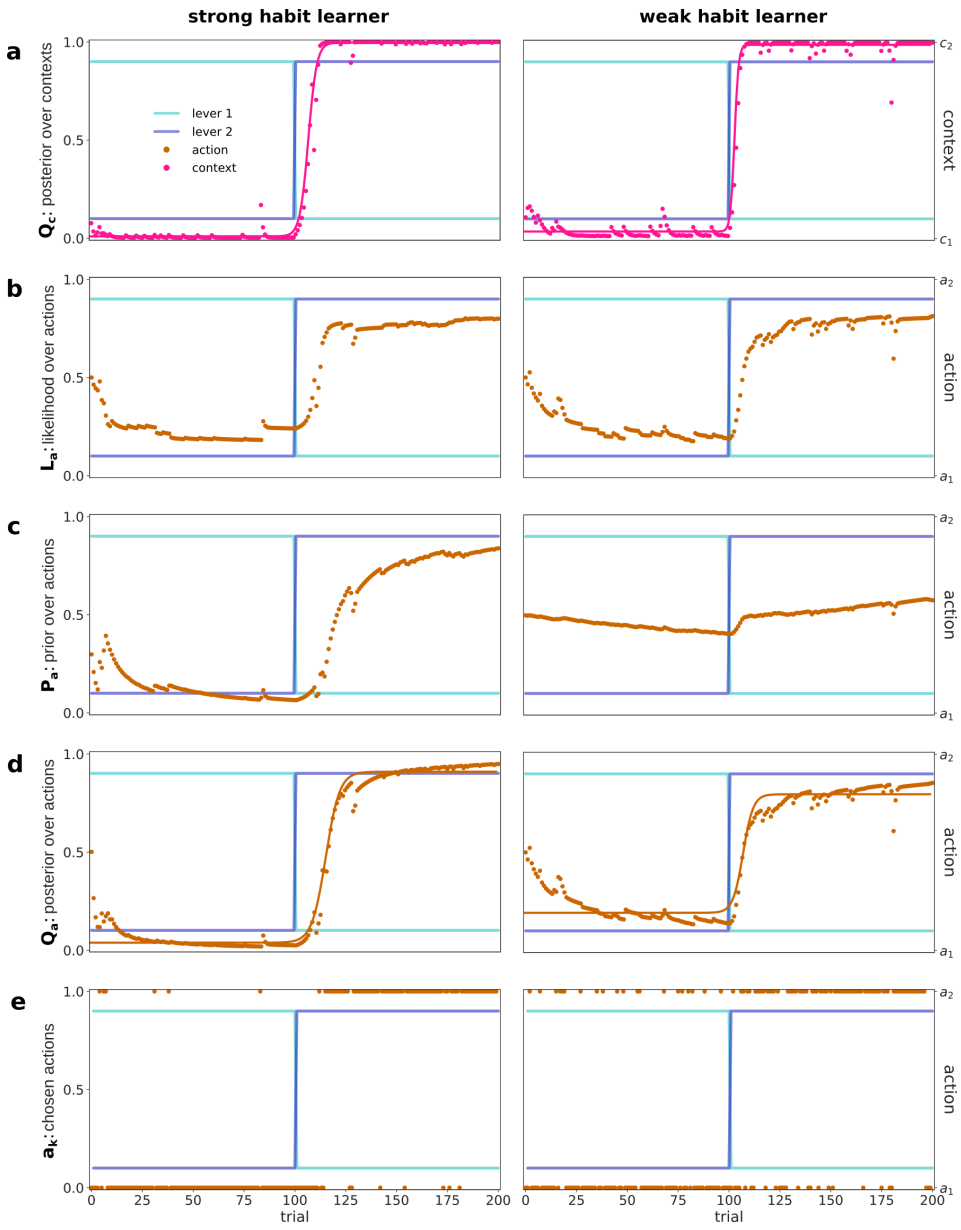

Figure 3.4: The dynamics of key internal variables of contextual habit learning agents during the habit learning task

The left column shows the dynamics for a strong ($h = 1.0$) habit learner and the right column for a weak ($h = 0.01$) habit learner. **a** The first row shows the agent's inference, the posterior beliefs over contexts $Q_{a'}$ i.e. the estimated probability of being in context 2. The pink dots are the agents' posterior beliefs in each trial of the task. The pink solid line is a fitted sigmoid, where its inclination point γ_3^c indicates when the posterior changes from representing context 1 to context 2. The light and dark blue lines are the reward probabilities of levers 1 and 2, respectively (see Figure 3.3). **b** The brown dots in the second row show the (normalized) likelihood L_a over actions. The likelihood encodes the goal-directed, anticipated value of actions, given the learned outcome contingencies. **c** The brown dots in the third row show the prior over actions $P_{a'}$ which encodes how likely the agent is a priori to select lever 2 and is a representation of the agent's habit. **d** The fourth row shows the posterior over actions $Q_{a'}$ which is the product of the prior and the likelihood. The brown dots show the posterior in each trial of the task, and the brown solid line shows a fitted sigmoid, whose inclination point can be interpreted as the trial at which an agent adapts its actions (see Figure 3.3d). **e** The brown dots in the bottom row show the chosen actions, which were sampled from the posterior over actions.

infer that the previously inferred context 1 is no longer an appropriate representation of the environment (see Figure 3.4a). Instead, the agents switch to believing to be in a new (second) context, and learn reward contingencies and habits for this context. The weak habit learner infers the context switch slightly earlier than the strong habit learner, at trials 103 and 107, respectively. According to the proposed model, the agents' context inference not only depends on surprising outcomes but also on the agents' own actions (see Section 3.3.4). The strong habit learner behaves highly consistently, even after the switch, and therefore is delayed in its context inference, relative to the weak habit learner. Note that the time point of this switch in beliefs was measured as the inflection point of a sigmoid fitted to the beliefs over time (a; solid line), see 3.3.5 and Figure 3.3d for a detailed explanation of how we used the parameters of the sigmoid.

Following context inference, the agents learn the new reward contingencies (see Figure 3.4b) and new habits (see Figure 3.4c) for context 2. Since this learning takes place after the context inference step, the posterior over policies is updated with a delay with respect to the context inference. As the agents sample their actions from the posterior, we can measure the trial at which they adapt their actions to press mostly lever 2 as the inflection point of the posterior. As with the posterior over contexts, we fitted a sigmoid (solid lines in Figure 3.4d) to calculate the time point of action adaptation, see Section 3.3.5 and Figure 3.3d.

In the following, we will call the time point (in trials) of action adaptation after the contingency change the habit strength, see 3.3.5. A value of 1 corresponds to the lowest possible habit strength, while a value of 100 means that an agent completely failed to adapt its behavior. This quantification is in line with the animal literature, where the amount of habitual behavioral control is measured by how often animals continue to choose the previously reinforced action after contingency degradation. As expected, the strong habit learner adapts its behavior later than the weak habit learner, at trials 116 and 107, respectively. This means the strong habit learner has a habit strength of 16 and the weak habit learner of 7.

The actions after the contingency switch in Figure 3.4e reflect this quantification of habit strength. The strong habit learner continues to choose lever 1 for around 10 trials, before it adapts and mostly consistently chooses lever 2 after 20 trials. The weak habit learner adapts earlier, but behaves less consistently and requires a longer transition period where both actions are chosen. However, due to the faster adaptation, in the first 15 trials after the switch, the weak habit learner exhibits a higher performance (chooses lever 2 in 47% of trials) than the strong habit learner (7% of trials, $p = 0.012$, two sample t-test on the actions in the first 15 trials after the switch).

The strong habit learner is able to recover its performance in the remainder of the extinction phase, where the task context is once again stable. Here, it not only learns the new reward contingencies, but a strong prior for action 2 (Figure 3.4c), so that it is again able to choose lever 2 more consistently, relative to the weak habit learner (92% vs 78%, $p = 0.01$, two sample t-test on the actions in trials $116 - 200$).

In summary, we found that a more pronounced prior causes as a stronger habit, as measured by the number of trial in the extinction phase before behavior is adapted. Critically, the mechanism is that a strong prior (Figure 3.4c) increases the certainty in the agent's posterior over actions (Figure 3.4d) and thereby its selection of the action (Figure 3.4e) with the higher expected reward. We found that as long as the environment is stable, the strong habit learner chooses the more rewarding option more reliably. This is the case in the training phase until the switch, and – after a brief adaptation period – after the switch. The strong habit learner exhibits less optimal behavior, in terms of obtained reward and relative to the weak habit learner, only immediately after the switch. This indicates that being a strong habit learner is

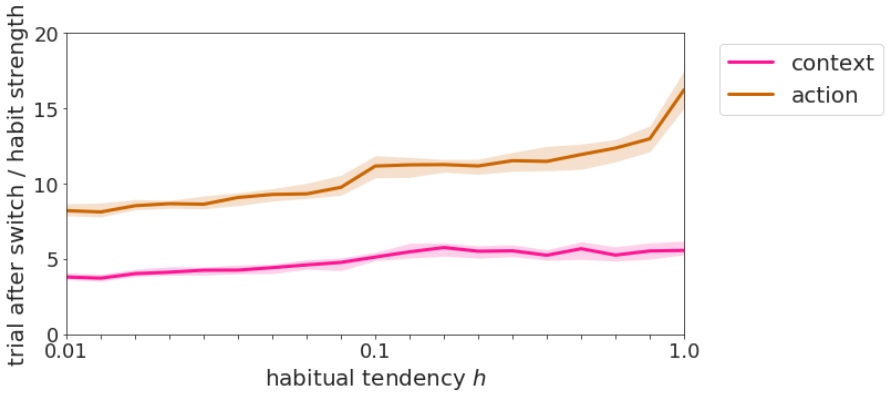

Figure 3.5: Habit strength as a function of the habitual tendency
For values of habitual tendency between 0.01 and 1.0, we plot the time points (in trials) of an inferred switch in context (pink solid line) and the habit strength (brown solid line). We measure habit strength as the time point of action adaptation after the switch, see Methods. For each habitual tendency value, we plot the median of 200 simulated runs, where the shaded areas represent the confidence interval of 95% around the median. We found a significant correlation between habitual tendency and habit strength ($p < 0.001$) and between habitual tendency and context inference ($p = 0.01$). The x-axis is logarithmically scaled.

useful for an agent, as long as contexts do not switch too often.

In addition, note that the effect of an increased certainty in action selection caused by the prior over actions is similar to a dynamic adjustment in decision temperature. Here, we did not use a decision temperature in our decision rule, as would be usually done in modeling noisy behavior (of participants), see Methods. Rather, we let the influence of the prior take this role. In the proposed context-specific model, this seems well motivated as the prior is learned conjointly with the reward contingencies, and indirectly reflects which behaviors have been successful in the past. This means that, in the proposed model, learned habits express themselves not only as an a priori preference for an action, but also as a dynamic adjustment of a decision temperature.

3.4.3 Habitual tendency increases habit strength

To generalize the effect of the habitual tendency on an agent's beliefs and behavior, we analysed agents with different values of the habitual tendency h, where we repeated simulations for each value 200 times, see Figure 3.5. The results confirm the conclusions drawn in the previous section: (i) All agents, independent of habitual tendency infer the context change quickly (within the first 5 trials after the switch), where strong habit learners infer the switch slightly later ($p = 0.01$, linear regression on the median values). (ii) Behavioral adaptation is at least 5 trials delayed compared to context switch inference. We find that acquired habit strength increases with the habitual tendency of an agent ($p < 0.001$, linear regression on the median values).

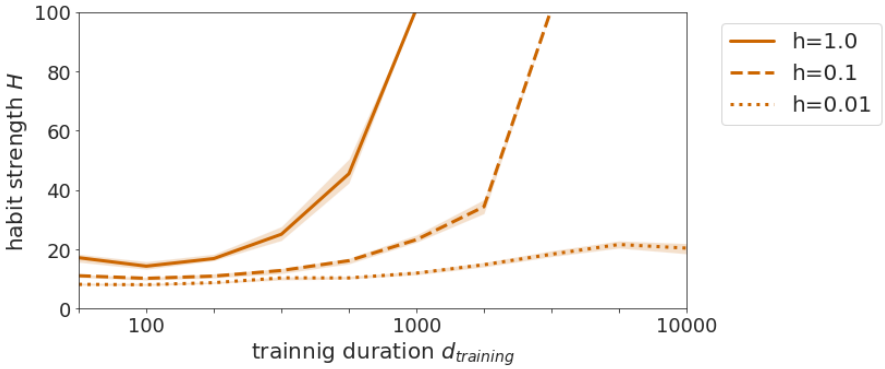

Figure 3.6: Habit strength as a function of training duration d_{training}
The x-axis is scaled logarithmically. The solid line represents a strong habit learner with a habitual tendency of $h = 1.0$, the dashed line a medium habit learner with $h = 0.1$, and the dotted line a weak habit learner with $h = 0.01$. The lines show the medians estimated over $n = 200$ repeated simulations for each level of the habitual tendency h. The shaded area shows the 95% confidence interval. A habit strength of 100 means that the posterior choice probability Q_a remains smaller than .5 during the entire 100 trials of the extinction phase.

3.4.4 Training duration increases habit strength

Here, we show that our proposed model is able to capture experimental findings that acquired habit strength depends on the amount of training a participant received. To test this, we simulated agents in the same habit learning task as above (see Figure 3.3) but now vary the length of the training phase d_{training} before the extinction phase.

In Figure 3.6 we plot the habit strength (see Methods) for three representative agents with different habitual tendencies (strong ($h = 1.0$), medium ($h = 0.1$), weak ($h = 0.01$)) as a function of training duration. For moderate training period durations ($d_{\text{training}} \leq 100$ trials), agents develop a relatively low habit strength and adapt their behavior rather quickly, within 20 trials. Although the differences are small for moderate training lengths, we find, as in the previous section, a significant correlation between habit strength and habitual tendency ($p < 0.001$, linear regression on the median values).

For longer training durations, habit strength is generally increasing. For very long training durations, both the strong and medium habit learner fail to adapt their behavior within the extinction period of 100 trials. The strong habit learner cannot adapt for a training duration $d_{\text{training}} \geq 1000$, and the medium habit learner for a training duration greater 5000. The weak habit learner exhibits only a slight increase in habit strength as a function of training duration.

In summary, these results stress the role of learning a prior over actions, where we interpret a strong prior as the representation of a habit, see e.g. Figure 3.4d. The longer the training period, the more pronounced the prior of a specific action will be, while the likelihood stabilizes after contingencies have been learnt properly (around 40 trials). Therefore the prior's influence on context inference and action adaptation increases with longer training periods, so that agents choose the previously reinforced action longer and longer in the extinction phase. The exact training duration at which adaptation starts to be delayed and fail depends on an agent's individual habitual tendency, where a higher tendency leads to a fail in adaptation for shorter training periods. This is in line with the literature on moderate and extensive training,

where extensive training leads to increased habit strength (Seger & Spiering, 2011).

3.4.5 Retrieval of previously learned context-specific habits

So far, we have assessed how habits can be represented as a prior over policies, where this prior is learned in a context-specific fashion. Here, we show a specific advantage of this context-specificity: The agent can recognize a previously experienced context by the associated contingencies and retrieve its habit (i.e., prior over actions) and learned reward generation rules for this context (Bouton & Bolles, 1979). As the prior implements a tendency to repeat actions, and actions were chosen according to their usefulness (i.e., likelihood of being chosen, see Fig. 3.1), habits in the proposed model represent which behavior is advantageous in a specific context. Therefore, recognizing the context and reusing previously established priors corresponds to a retrieval of previously learned optimal behavior, i.e., habits.

In Figure 3.7, we show the design of the 'habit retrieval experiment', which is an extension habit learning task. As before, we first let agents experience the two contexts for 100 trials each, and call this the learning phase of the experiment. Critically, there is an, additional phase, the retrieval phase, where we place agents again into context 1 for 100 trials. In the first trial of this retrieval phase, we induce maximal uncertainty about the context by setting the agents' prior over contexts to $p\left(c_{201}\right) = (0.5, 0.5)^T$. Here, we wanted to emulate a situation where an agent knows there is a context change, but not to which context, akin to a mouse being taken out of its home cage into the experimental setup. If we had kept the prior over contexts as the old posterior from the last trial of the learning phase, we would induce habit effects where agents delay adaptation for the reasons discussed in the previous sections. The setup is similar to the experimental setup used in (Bouton & Bolles, 1979; Gershman et al., 2010). To compare 'experienced' agents with agents that have not learned yet context 1, we implement 'naive' agents, as in Section 3.4.2.

To quantify the advantage of the retrieval of previously learned context-specific behavior, we first measured how long it takes a naive agent to converge to a stable beliefs level about context 1 in the learning phase (Figure 3.8a). To evaluate the convergence times to a stable knowledge for naive agents, we fit again a sigmoid to the posterior over contexts and actions in the learning phase (as in Figure 3.4, see also Methods). We interpret the left asymptote $\gamma_4^{a,c}$ of the sigmoid as the stable level of knowledge the agents eventually reach. We calculate the convergence time as the trial in which the posterior crosses the left asymptote for the first time. We compare this duration to how long experienced agents take to recognize the known context 1 in the retrieval phase and reuse their previously learned behavior. To compute convergence times for the experienced agent, we determined the first trial in the retrieval phase where the posterior is lower than the left asymptote which was fitted for the learning phase.

These convergence times, as a function of habitual tendency, are shown in Figure 3.8 for both the naive and the experienced agents. We discussed the initial development and convergence of the posteriors shown in Figure 3.4 for single runs of agents in Section 3.4.2. The results here are a quantification of these for different habitual tendencies using 200 runs each. Naive agents (see Figure 3.8a) are able to achieve a stable level of knowledge for the context in around 8 trials, if they have a low habitual tendency (e.g. 0.02), and in around 5 trials, if they have a high habitual tendency (1.0). As discussed above, context convergence time are faster for higher habitual tendency, because these depend partially on the agent's

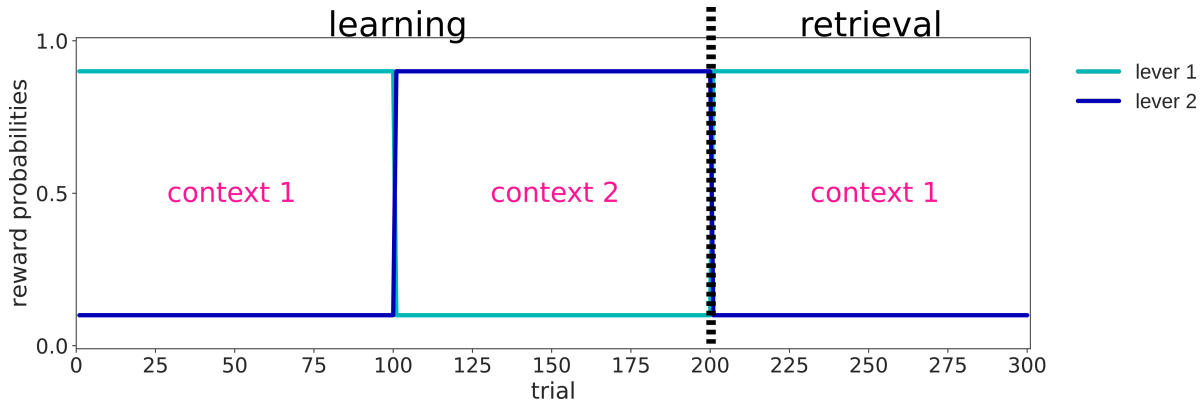

Figure 3.7: The habit retrieval experiment
A 300 trial experiment consisting of a learning phase (equivalent to the whole habit task, see Figure 3.3) with 200 trials, and a new, additional habit retrieval phase with 100 trials. The light blue line shows the probability of lever 1 paying out a reward, and the dark blue line shows the probability of lever 2 paying out a reward. The vertical dashed line indicates the switch from the learning to the retrieval phase. In the retrieval phase, the agent revisits context 1, where outcome contingencies are exactly the same as in the first 100 trials of the experiment.

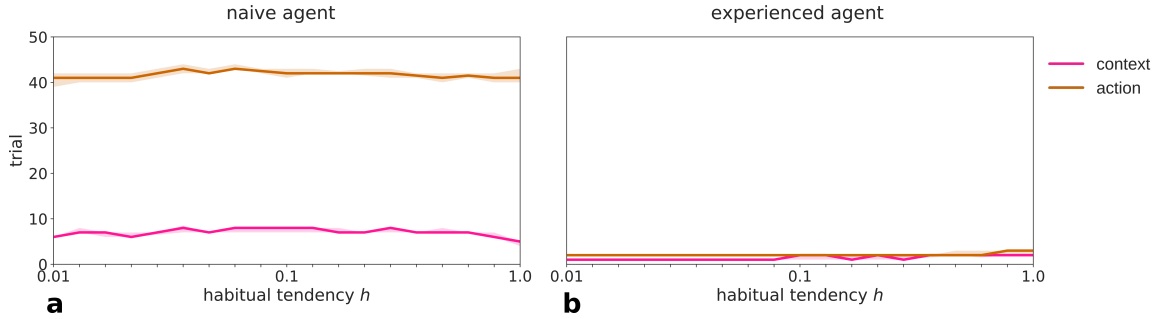

Figure 3.8: Convergence times of the posteriors as a function of the habitual tendency h
a Convergence times of the posterior beliefs over contexts (pink) and actions (brown) in naive agents who visit context 1 for the first time, see main text how convergence times were quantified. The shaded areas indicate a confidence interval of 95%. A naive agent takes around 7 trials to converge to stable beliefs about its context. It takes around 40 trials to converge to a stable posterior over actions, indicating the time it takes to learn a stable representation of the action-outcome contingencies for this context. **b** Convergence times of the posterior beliefs in naive agents who visit context 1 for the second time. An experienced agent takes 1 to 2 trials to recognize it is in the known context 1. It almost instantly retrieves its knowledge about outcome contingencies and its habit for this context, and thereby its posterior over actions, so that the action adaptation happens maximal one trial later.

own more consistent behavior. Action convergence times mainly depend on learning the outcome rules and the resulting likelihood, which takes, for the naive agent, with around $40 - 45$ trials a lot longer than context inference. We find that these times are not influenced by an agent's habitual tendency.

For experienced agents, both, recognition of the known context, as well as reusing the old outcome rules and habits, happens almost instantaneously, within first 3 trials of the retrieval phase, see Figure 3.8b. As a consequence of these faster convergence times, experienced agents choose the optimal lever more often in the retrieval phase than in the first half (context 1) of the learning phase (94% vs 87%, $p < 0.001$, two-sample t-test, averaged over all habitual tendencies). In addition, we find that agents continue to learn outcome contingencies and habits during the renewed exposure to context 1 (data not shown). Importantly, in terms of behavior, for both the naive and experienced agent, the percentage of choosing the optimal action increases with habitual tendency (naive: $p = 0.001$; experienced: $p = 0.036$; linear regression on the median values). This finding provides another hint that being a strong habit learner might be advantageous if one's environment is mostly stable except for sudden switches to already known contexts, see also Discussion.

3.4.6 Environmental stochasticity increases habit strength

In this section, we examine how environmental stochasticity, namely the probability of observing a reward, interacts with the habit learning process (DeRusso et al., 2010). We again let artificial agents perform in the habit learning task (see Figure 3.3). We varied the probability of receiving a reward ν in both the training and extinction phases from $\nu = 1.0$ (completely deterministic) to $\nu = 0.6$ (highly stochastic, where a 0.5 probability would mean that outcomes are purely random). In the extinction phase, lever 1 has probability ν to pay out a reward, while lever 2 pays out a reward with a probability of $1 - \nu$. These probabilities are reversed in the extinction phase.

Figure 3.9 shows the habit strength, measured in the extinction phase as a function of environmental stochasticity $1 - \nu$. As before, we used three agents with different habitual tendencies (strong ($h = 1.0$), medium ($h = 0.1$), weak ($h = 0.01$)). In a fully deterministic environment ($1 - \nu = 0$), all three agents have a similarly low habit strength (below 10). The agents infer the context switch immediately (not shown) and adapt their behavior shortly after. When the reward probability is $\nu = 0.9$ and the stochasticity is $1 - \nu = 0.1$, we find habit strengths between 7 and 15, which replicates the result shown in Figure 3.5. For more stochastic rewards, we find that for all three agents the habit strength increases with stochasticity, until they fail to adapt within the extinction phase. In addition, one can see that the habit strength is higher, the higher the habitual tendency of the agent is ($p < 0.03$ for a ANOVA on parameters of fitted exponential functions), and the exact amount of stochasticity agents can handle before they fail to adapt depends on the agent's habitual tendency.

In the model, this effect is due to two factors: First, as the environment becomes more stochastic, it is harder for an agent to detect the switch contingencies. This delays context inference and thereby action adaptation. Second, the likelihood encoding the goal-directed value is less pronounced in a stochastic environment, as it maps to the decreased probability of achieving a reward for an action. In the model, the agent samples actions from the posterior, which is the product of the likelihood and the prior. If the likelihood is less pronounced, the habits, as represented by the prior, will automatically gain more weight in the posterior, leading to an increased reliance on habitual behavior in a stochastic environment. Intuitively,

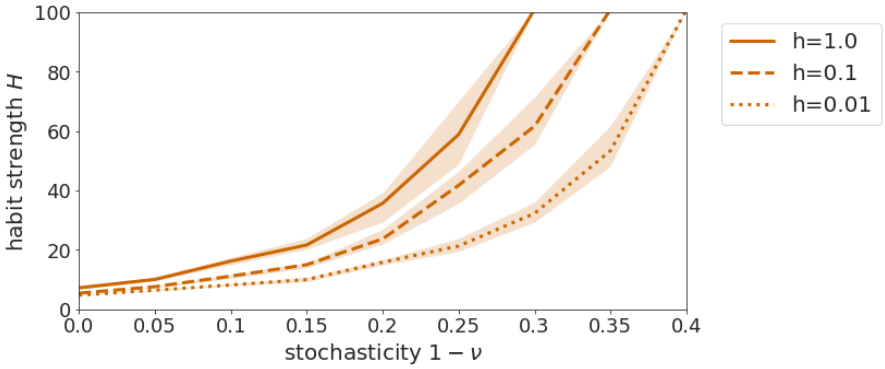

Figure 3.9: Habit strength as a function of environmental stochasticity $1 - \nu$
The three habit learners (strong, medium, weak) develop stronger habits if the reward scheme is more stochastic, i.e. reward probabilities ν are lower. Solid line: strong habit learner with $h = 1.0$; dashed line: medium habit learner with $h = 0.1$; dotted line: weak habit learner with $h = 0.01$. The shaded area surrounding the lines is the confidence interval of 95%. A habit strength of 100 means that the agent does not adapt its behavior within the extinction period of 100 trials.

this means that a decrease in goal-directed value of actions gives way to a stronger influence of habits. Conversely, habits are also learned more slowly in more stochastic environments because actions are not chosen as consistently because of the decreased goal-directed value. We will come back to the important implications of these findings in the Discussion.

3.4.7 Outcome devaluation

In this final results section, we show that the proposed model can also qualitatively replicate results from outcome devaluation studies, e.g. (Adams, 1982). We modified the habit learning task (Figure 3.3) by introducing an outcome devaluation in the extinction phase, in addition to the switch in outcome rules. This was done by reducing the prior preference for the reward of lever 1 but not lever 2 in the extinction phase (for details see Appendix).

In general, we find that the outcome devaluation results in a discontinuous jump in the likelihood, as the devalued reward means that action 1 suddenly has no more goal-directed value (data not shown) while action 2 remains useful. Nonetheless, we can apply the same analyses as in Section 3.4.3 to show the effect of habitual tendency on habit strength under outcome devaluation.

Figure 3.10 shows, as a function of habitual tendency, (i) the trials numbers in the extinction phase when agents inferred a switch in context and (ii) habit strengths. Independent of habitual tendency ($p = 0.54$, linear regression), agents infer the context switch slightly earlier than in the task without outcome devaluation (median trials 2.4 vs 3.6, $p < 0.001$, two-sample t-test). As before, agents with a low habitual tendency (≤ 0.02) only develop a very weak habit within the training phase of 100 trials (see Figure 3.4c). When the usefulness of actions now changes due to the devaluation, these agents can instantly, at the beginning of the extinction phase, adapt their behavior to start pressing lever 2. Agents with a higher habitual tendency (≥ 0.1) on the other hand, learn a pronounced habit during training. As a result, these strong habit learners show in the extinction phase after devaluation a delayed action adaptation and

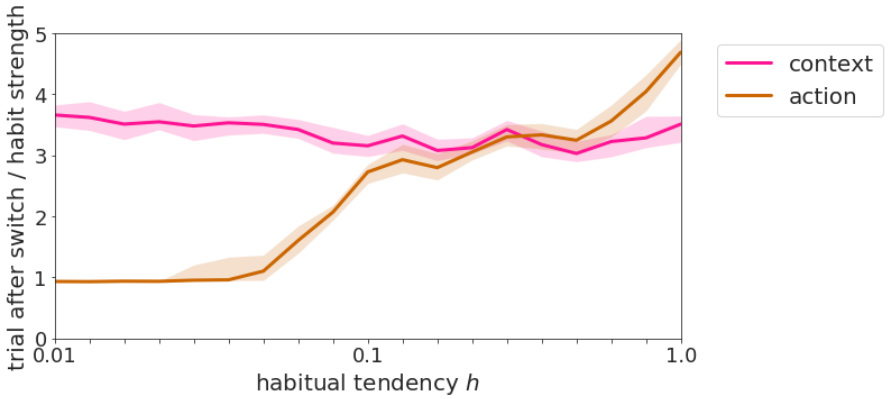

Figure 3.10: Context inference and action adaptation in the devaluation experiment
Pink line: Time points in the extinction period of when agents infer a switch in context, as a function of the habitual tendency. Brown line: habit strength, as a function of the habitual tendency. The x-axis is logarithmically scaled. This figure is based on the same analysis methods as Figure 3.5, but here we analyzed the posteriors in an environment with contingency degradation and outcome devaluation.

thereby a habit strength greater 1 (up to 4). Generally, as before, a higher habitual tendency leads to a greater habit strength ($p < 0.001$, linear regression on the medians).

Clearly, we found a devaluation effect for agents with a habitual tendency $h \geq 0.1$. Although the habit strengths are fairly low, we found that if we increase the training duration to more extensive training (≥ 500 trials), habit strength increases, so that even weak habit learners show a habit strength greater than 1, and strong habit learners have a habit strength of up to 8 (data not shown).

While these effects are lower than in the contingency degradation experiment, these results show that our model can in principle emulate habitual behavior in both classical experimental designs, contingency degradation and outcome devaluation (Yin & Knowlton, 2006).

3.5 Discussion

In this paper, we proposed a Bayesian contextual habit learning model. In this model, habits are the prior over policies, which implements an a priori and value-free bias to repeat previous policies, while the goal-directed evaluation is represented by a likelihood, which encodes the anticipated goal-directed value of policies. An agent who uses this model for action selection samples actions from the posterior, which is the product of the prior times the likelihood. One of the key results is that this rather simple procedure implements an adaptive and automatic balancing of goal-directed and habitual control. An important ingredient for this procedure to work is that habits and outcome rules are learned in a context-specific manner so that an agent can learn and retrieve specific habits and outcome rules for each context it encounters. We used a free (adjustable) parameter to model a trait-like habitual tendency h, which determines the learning rate of the prior over policies, and thereby the acquisition speed of the habit. We introduced a habit learning task with a training and extinction phase, and showed the basic properties of an agent's information processing employing the model. Using agent-based

simulated experiments, we were able to show that our model captures important properties of experimentally established habit learning: insensitivity to both contingency degradation and outcome devaluation, increased habit strength both with extended training duration and with increased environmental stochasticity, and near-instantaneous recovery of habits when exposed to a previously experienced context. We also found that the habitual tendency interacts with these effects: Agents with higher habitual tendencies exhibit increased habitual contributions to control and habit strength in all of these experimental conditions.

In recent years, several approaches to computationally model goal-directed and habitual behavior have been proposed. An often used interpretation of two distinct habitual and goal-directed systems has been the mapping to model-free and model-based reinforcement learning (Daw et al., 2011). Here, the model-free system implements an action evaluation based on which actions have been rewarding in the past. The model-based system implements goal-directed forward planning resting on a Markov decision process. Typically, these models have to be run in parallel and require an additional arbitration unit, which evaluates both systems and assigns a weight to each, determining the respective influence on action selection, see e.g. (S. W. Lee et al., 2014). However, it seems an open question, whether model-free learning can be indeed mapped to habitual control. For example, (Friedel et al., 2014) were able to map model-based reinforcement learning to goal-directed behavior but failed to find such a relation for habitual behavior and model-free reinforcement learning, see also (Wood & Rünger, 2016) for a recent review about the relationship between habitual control and model-free learning.

To resolve this issue, (Miller, Ludvig, Pezzulo, & Shenhav, 2018) proposed to map habitual behavior to a value-free system, which implements a tendency to repeat actions. In this view, the goal-directed system corresponds to a value-based system, which includes model-based as well as model-free reinforcement learning, and both systems are arbitrated using an additional arbitration unit. Our model aligns with this proposal, as we model the prior as based on pseudo-counts which indicate how often an action has been chosen in the past. As a result, an action will have a higher habitual weight if it has been chosen more often, implementing a habit based on repetition of previous behavior. Goal-directed control is described based on a Markov decision process as well, which in our model is solved using Bayesian methods, instead of reinforcement learning. Despite these conceptual similarities with regard to the interpretation of the nature of habitual behavior, the proposed value-free model is fairly different from the model presented here. A key difference is that we used a hierarchical model to implement context-dependent learning, which we found essential for reproducing key features of habitual behavior. Furthermore, our model does not require an additional arbitration unit with additional computational costs. Rather, in the present model, habitual and goal-directed contributions are balanced directly using Bayes rule. In other words, we interpret experimental evidence for habitual and goal-directed control not as evidence for a dichotomy that competes for action control. Rather, we see action control as a probabilistic inference problem, where two sources of information are integrated: The likelihood which looks at the situation at hand, and the prior which is shaped by past experience.

There have been other Bayesian proposals to habit learning, particularly using active inference. (FitzGerald et al., 2014a) and (K. Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty, & Pezzulo, 2016) regarded habits in a similar manner to model-free learning, and implemented them as an additional simplified policy. This approach is therefor fundamentally different from and potentially complementary to ours. Nonetheless, we think it possible that the brain uses both value-free as well as model-free learning processes, so that it would not be unreasonable to assume that both contribute to action selection. (Maisto et al., 2019) regarded

habits as cached values of the likelihood calculated in previous trials of the experiment. This means that the likelihood was only calculated when first encountering a new context, and is then kept stable and cached as long as the context does not change. These proposals of a Bayesian treatment of habit learning are different from (and possibly complementary to) our approach, as we view habits as a prior over actions or policies, and not related to the likelihood (which in our model represents goal-directed control). Under extensive training regimes, both approaches might lead to similar results. However, under limited training, when both, goal-directed and habitual control influence behavioral control, our approach may lead to more plausible behavior in this regime because of the balancing of the two contributions.

Furthermore, there are other proposals to view reinforcement, contingency degradation, and outcome devaluation experiments as a context inference and rule learning problem (Palminteri et al., 2015; Gershman et al., 2010; Wilson et al., 2014). These studies view task states as latent variables or contexts, which need to be inferred, while reward generation rules from these states are learned, which essentially translates to a non-hierarchical, partially observable Markov decision process. What sets our proposal apart, is that we view the context as a latent variable on a higher level of a hierarchical model, which modulates how rewards are generated from the same states. This allows us to describe not only actions but sequences of actions which enables an agent to navigate a state space, where the rules might change even within the same environment. We can thereby incorporate the assumption that habits are based on chunked action sequences which allows us to map our model to interesting neurophysiological findings which we discuss below. These ideas also align with proposals such as event coding (Hommel et al., 2001) and event segmentation theory (Zacks et al., 2007), which posit that behavior is segmented into events or episodes. Based on these proposals, (Butz, Bilkey, Humaidan, Knott, & Otte, 2019) put forward an interesting context and contingency learning model which implements ideas similar to our goal-directed evaluation in a neural network model.

Typically habit experiments focus on providing evidence that animals have acquired a habit, which are by experimental design non-functional, as the habit is measured by its suboptimality, i.e. to perform an action even though it will not longer produce a reward. In contrast, as we have access to the internal variables of the agent, we can observe, in addition, subtle changes in behavior and the causes for these changes before and during extinction and devaluation. This perspective allows us to asses the advantages of being a strong habit learner, e.g. (Wood & Rünger, 2016): Fast and efficient action evaluation, choosing consistent and reliable behavior, especially in uncertain conditions, an increased success rate, and quick retrieval of previous habits in a known context which amounts to retrieving optimized behavior. In the following, we will discuss how these advantages come about in terms of the proposed model.

According to the model, habits are fast and efficient because the prior over policies is retrieved from some context-specific 'prior over policies memory' and is not evaluated in a costly manner. Interestingly, we found that being a strong habit learner supports choosing consistent, reliable behavior. For example, agents with higher habitual tendencies chose the better option more reliably in our tasks. This was true as long as agents were in a stable context where outcome rules did not change. Only in the short time after a contingency switch did they choose the unrewarded option more often due their delayed behavioral adaptation, in comparison to an agent with a low habitual tendency. Strikingly, precisely this effect has been observed in a recent study, where (McKim, Bauer, & Boettiger, 2016) found that participants with a history of substance use disorder (SUD) have a heightened ability to execute previously learned stimulus-response associations, in comparison to controls. Assuming that a history

of SUD is correlated with a stronger tendency to learn habits, this result directly reflects on our finding of an increased performance for higher habitual tendencies in known contexts. As we found in our simulated experiments, the participants with presumed higher habitual tendency (SUD history) were also found to show a decrease in performance after a switch to a new context, and showed signs of perseverance of behavior.

This effect of improved choice behavior was also seen when agents revisited a known context. Here, the already learned, contextual habit enables an agent to quickly retrieve previously acquired behavioral patterns for this context, which are presumably optimal if the contingencies of the context did not change between the two visits. Importantly, being a strong habit learner also helped performance in uncertain conditions: When rewarding outcomes were highly stochastic, we found that the habit (prior over policies) has a stronger weight on action selection and helps an agent choosing the better option more reliably. Taken together this means that being a strong habit learner is advantageous, as long as one's environment is subdivided into stable phases of already known contexts, separated by infrequent switches. Interestingly, there is evidence for such a mechanism of rapid context-dependent habit retrieval (Bouton & Bolles, 1979; Gershman et al., 2010): Using optogenetics, (Smith, Virkud, Deisseroth, & Graybiel, 2012) observed rapid re-instantiations of a previously learned habit after a context change, where, similar to our simulated experiments, reward contingencies changed. Obviously, this advantage of habits may be even increased, if agents, as is the case in our real-life environment, were able to choose the context they are in or switch to. While we did not implement this active component here, it would most likely lead to agents choosing long stable contexts for which they already learned habits. These scenarios would lead to interesting research about how agents decide to switch contexts to balance exploration and exploitation in their environmental niche.

We identified several causes of variability in habitual control in the agent. First, as habits in the form of a prior over actions are learned by exposure to stochastic stimuli, their contribution is therefore dynamic and adaptive during a task. In other words, in our model, an agent never stops adapting a habit so that habit strength varies and is context- and experience-dependent. Secondly, we found that behavior is strongly controlled by habits in those situations when goal-directed forward planning cannot determine a clearly best action, so that there is uncertainty on what the best course of action is. This means that habits, when there is conflict between different possible (goal-directed) actions, can be seen as an informed guess to select an action and resolve the conflict rapidly. This uncertainty-weighting of control is in line with previous findings (Daw et al., 2005; S. W. Lee et al., 2014). Thirdly, we found that one can emulate an individual habitual tendency simply by varying the initial pseudo counts of the prior so that the individual learning rate during habit acquisition varies, which in turn leads to variations in delayed action adaptation and different habit strengths.

Even though there are advantages of using habits, it clearly depends on the environment whether habits will be mostly advantageous or disadvantageous. For example, a strong habit learner would be best placed in an environment with rare switches between already learned contexts, see e.g. (Barnes, Kubota, Hu, Jin, & Graybiel, 2005; Gremel & Costa, 2013). Conversely, an environment with frequent changes between contexts dissimilar from previously learned ones would lead to decreased choice performance of a strong habit learner, in comparison to a weak habit learner. Another possibility how the habitual control mechanism may be detrimental to performance is if context inference is for some reason dysfunctional. For example, with suboptimal context inference, one may expect that there is confusion between contexts that are similar in appearance but effectively distinct. We speculate that this confusion may express itself experimentally as an apparent decrease of top-down control by cortical

areas (context-inference) on the striatum (habitual control), as e.g. found in (Renteria, Baltz, & Gremel, 2018). Another interesting and experimentally relevant example of biased context inference may be the established phenomena of Pavlovian to Instrumental Transfer (PIT), (Garbusow et al., 2014; Talmi, Seymour, Dayan, & Dolan, 2008), where participants are biased towards a previously encountered context by cues of that context. Note that in our model, contexts are not cued and instead need to be implicitly inferred from the observed reward rules of the environment, where we refer to a context as a specific set of states and their corresponding outcome rules. Nonetheless, even without cues, retrieval of previously learned habits was almost instantaneous, which would only be facilitated if, in addition, cues were presented.

This mechanism may relate to addictive behavior like substance use disorders (SUD), which are characterized by a shift from goal-directed to habitual and compulsive use. We speculate that difficulties in context inference may help explain how addictive behavior becomes habitual: While there is a clear difference in the outcomes between initial substance use (euphoria or relaxation) and the outcome after a prolonged time period of use (e.g. adverse health or social consequences), the user may not infer that these two outcomes are two different contexts. Additionally, the use is typically associated with some stimuli or cues, like the ringing sound of glasses, which become connected with the context and associated contingencies. As outcomes become gradually less rewarding, the cues remain the same, and the contingencies may not change quickly enough to be sufficiently driving a change in context inference. Consequently, the action control of the user might not infer that prolonged use has placed the user into a qualitatively new context, in which the initially learned habit provides for suboptimal behavior. With suboptimal context inference in place, behavior will be strongly biased by the already learned habit. As habits are hard to unlearn within a context, the user will have difficulties to unlearn the habit. As uncertain probabilistic rewards shift control further to habits, the difficulty to unlearn is further enhanced, where the reward stochasticity may result from differences of outcomes but also from the user's memory of the desirable outcomes after initial substance use. It is an open question, whether people who become addicted have a higher habitual tendency, or/and whether drugs of abuse increase an individual's habitual tendency. Another potential reason in the model that action control will be biased towards habits is if the likelihood, i.e. the goal-directed evaluation, does not produce a clearly best action, e.g. due to uncertainty about goals or a relatively low planning depth. According to the model, limited planning capacity would translate into a less accurate and potentially less pronounced likelihood, which leads to the habitual prior automatically gaining more weight in the action selection. This holds while learning habits, but also when reentering a known context.

Although we have used in our simulated experiments policies of just a single action ($len\,(\pi) = 1$), the proposed model also supports behavioral episodes and policies with length $len\,(\pi) > 1$, i.e. sequences of actions. Interestingly, a growing and compelling area of research is to view habits as chunked and automatic action sequences (Graybiel & Grafton, 2015; Smith & Graybiel, 2016), which might be embedded in a hierarchical model (Dezfouli & Balleine, 2012, 2013). This sequential view on habits rests on both neurophysiological and behavioral evidence, see (Smith & Graybiel, 2014; Corbit, 2018) for recent reviews: In animal experiments, both the dorsolateral striatum (DLS) and infralimbic (IF) cortex have been found to be implicated in habitual control and to exhibit so-called (task-) bracketing activity, where neurons are active at the beginning and the end of an action sequence, e.g. (Smith & Graybiel, 2013). The computational function of this bracketing activity is yet unclear. We speculate, building on insights from the proposed model, that this bracketing activity may be the expression of

a context-dependent prior over policies being set at the beginning of an action sequence, see Fig. 3.1. Setting such a prior has the advantage that the organism, during executing a fast, but controlled action sequence, can focus only on a single or few policies whose prior is greater than 0. This focus enables fast action control as only for these few policies the likelihood needs to be evaluated. Critically, in the proposed model, the computation of prior and the likelihood of policies have a clear sequential order in time; as the prior refers to what policies in a specific context are predicted to be useful, even before the organism has actually evaluated any policy, selecting this prior clearly has temporal precedence, as in the proposed model, over the evaluation of the likelihood during performing the action sequence. Precisely this temporal precedence has been observed experimentally during habit learning: First, the beginning of the bracketing activity, e.g. in DLS, could be interpreted as a retrieval and encoding of a prior over policies, while subsequent activity during the action sequence, e.g. observed in dorsomedial striatum (DMS), could be an expression of the evaluation of the likelihood over policies and the computation of a posterior over actions, i.e. once the organism is receiving sensory input caused by executing the selected policy. Experimentally, this DMS activity has been reported to be mostly present during rather early stages of habit learning, and to decrease over time until a habit has been learned (Thorn, Atallah, Howe, & Graybiel, 2010). In our model, this gradual decrease, over time, of DMS activity is reflected by the increasing prior and posterior over policies over trials, e.g., see Figure 3.4c,d. Finally, the bracketing activity in DLS at the end of an action sequence can be explained in the proposed model by the updating of the prior over policies after performing the action sequence (see Fig. 3.1), in particular the 'sharpening' of end activity and the reduction in entropy, over trials, as reported in (Desrochers, Amemori, & Graybiel, 2015). Findings of lesioning experiments also fit into this picture: After habit learning, lesioning DLS led to a behavioral switch from habits back to goal-directed action while lesioning of the DMS had no apparent effect (Yin et al., 2004). Similarly, in another study, at an early stage of habit learning, inactivation of DMS reduced the goal-directed response (i.e., in the model, likelihood and posterior can no longer be computed) while inactivation of DLS was without effect (i.e., in the model no prior over actions had be learned yet) (Corbit, Nie, & Janak, 2012). In the same study, after habit learning, inactivation of DLS let the animals return to goal-directed behavior (i.e., in the model, the prior over policies is now flat) while inactivation of DMS is without effect (i.e., in the model, the prior over policies outweighs the now flat likelihood). Future studies will have to experimentally test whether our predictions hold, and if our model indeed maps to these brain structures.

In summary, the proposed modelling approach provides the novel perspective that habitual control relies on learned context-specific priors of policies. The resulting model provides for a simple way to balance action control between habits and goal-directed control. As we have discussed, experimental findings seem to support this perspective of a separation into prior and posterior over policies. We anticipate that the present computational modelling approach may support novel directions of research aimed at the central role of context inference as a means to reduce the number of policies that have to be evaluated and implement fast action control relying on the interplay between the prior and posterior over policies.

3.6 Acknowledgments

We thank Ann-Kathrin Stock for valuable comments and suggestions.

3.7 Funding acknowledgments

Funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft), SFB 940/2, projects A9 and Z2, and TRR 265, project B09 (SJK).

3.8 Appendix

3.8.1 Derivations of the update equations

The variational free energy functional is defined as the Kullback-Leibler divergence between the approximate posterior 3.6 and the joint probability distribution of the generative model 3.4. Hence, we can write the variational free energy as

$$
\begin{aligned}
F\left[q\right] =& D_{KL}\left[q\left(\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T}, \pi, \theta, \phi, \mathbf{c}_k\right) | p\left(\mathbf{s}_{1:T}, \mathbf{r}_{1:T}, \pi, \theta, \phi, \mathbf{c}_k\right)\right] \\
=& \sum_{\mathbf{c}_k} q\left(\mathbf{c}_k\right) \left[\ln \frac{q\left(\mathbf{c}_k\right)}{p\left(\mathbf{c}_k\right)} + \int d\phi q\left(\phi\right) \left[\ln \frac{q\left(\phi\right)}{p\left(\phi\right)} + d\phi q\left(\theta\right) \left[\ln \frac{q\left(\theta\right)}{p\left(\theta\right)} + \sum_{\pi} q\left(\pi|\mathbf{c}_k\right) \left[\ln \frac{q\left(\pi|\mathbf{c}_k\right)}{p\left(\pi|\theta, \mathbf{c}_k\right)} \right.\right.\right.\right. \\
& \left.\left.\left.\left. - \ln p\left(\mathbf{s}_{1:t}, \mathbf{r}_{1:t}|\pi, \phi, \mathbf{c}_k\right) + \sum_{\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T}} q\left(\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T}|\pi, \mathbf{c}_k\right) \ln \frac{q\left(\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T}|\pi, \mathbf{c}_k\right)}{p\left(\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T}|\mathbf{s}_t, \pi, \phi, \mathbf{c}_k\right)} \right]\right]\right]\right]
\end{aligned}
\tag{3.15}
$$

where for clarity we omitted the parametric dependence of each distribution. The approximate posterior is than obtained as the minimum of the free energy, defining the upper bound on surprise (negative marginal log-likelihood).

We first write down the update equations for the beliefs over future states and rewards within an episode, using the belief propagation message passing update rules (Pearl, 2014; Yedidia et al., 2003b). For details on the derivation steps see our previous work (Schwöbel et al., 2018) in which we investigated the Bethe approximation for a Bayesian treatment of a partially observable Markov decision process. The results shown here are an adaptation for fully observable states, which is just a special case.

$$
\begin{aligned}
q\left(\mathbf{r}_\tau, \mathbf{s}_\tau|\pi, \mathbf{c}_k\right) &= \frac{p\left(R=1|\mathbf{r}_\tau\right) p'\left(\mathbf{r}_\tau|\mathbf{s}_\tau, \mathbf{c}_k\right) m_\pi^{\tau+1}\left(\mathbf{s}_\tau|\mathbf{c}_k\right) m_\pi^{\tau-1}\left(\mathbf{s}_\tau|\mathbf{c}_k\right)}{Z_\tau^\pi} \\
q\left(\mathbf{r}_\tau|\pi, \mathbf{c}_k\right) &= \frac{p\left(R=1|\mathbf{r}_\tau\right) m_\pi^\tau\left(\mathbf{r}_\tau|\mathbf{c}_k\right)}{Z_\tau^\pi} \\
q\left(\mathbf{s}_\tau, \mathbf{s}_{\tau-1}|\pi, \mathbf{c}_k\right) &= \frac{p\left(\mathbf{s}_\tau|\mathbf{s}_{\tau-1}, \pi\right)}{Z_{\tau,\tau-1}^\pi} m_r^{\tau-1}\left(\mathbf{s}_{\tau-1}\right) m_r^\tau\left(\mathbf{s}_\tau\right) m_\pi^{\tau+1}\left(\mathbf{s}_\tau|\mathbf{c}_k\right) m_\pi^{k-2}\left(\mathbf{s}_{\tau-1}|\mathbf{c}_k\right) \\
q\left(\mathbf{s}_\tau|\pi, \mathbf{c}_k\right) &= \frac{m_\pi^{\tau+1}\left(\mathbf{s}_\tau|\mathbf{c}_k\right) m_\pi^{\tau-1}\left(\mathbf{s}_\tau|\mathbf{c}_k\right)}{Z_\tau^\pi}
\end{aligned}
\tag{3.16}
$$

using the messages

$$
\begin{aligned}
m_r^\tau\left(\mathbf{s}_\tau|\mathbf{c}_k\right) &= \sum_{\mathbf{r}_\tau} p\left(R=1|\mathbf{r}_\tau\right) p'\left(\mathbf{r}_\tau|\mathbf{s}_\tau,\mathbf{c}_k\right), \\
m_\pi^\tau\left(\mathbf{r}_\tau|\mathbf{c}_k\right) &= \sum_{\mathbf{s}_\tau} p'\left(\mathbf{r}_\tau|\mathbf{s}_\tau,\mathbf{c}_k\right) m_\pi^{\tau+1}\left(\mathbf{s}_\tau|\mathbf{c}_k\right) m_\pi^{\tau-1}\left(\mathbf{s}_\tau|\mathbf{c}_k\right), \\
m_\pi^{\tau+1}\left(\mathbf{s}_\tau|\mathbf{c}_k\right) &= \frac{1}{Z'_{\tau,\pi}} \sum_{\mathbf{s}_{\tau+1}} p\left(\mathbf{s}_{\tau+1}|\mathbf{s}_\tau,\pi\right) m_r^{\tau+1}\left(\mathbf{s}_{\tau+1}|\mathbf{c}_k\right) m_\pi^{\tau+2}\left(\mathbf{s}_{\tau+1}|\mathbf{c}_k\right), \\
m_\pi^{\tau-1}\left(\mathbf{s}_\tau|\mathbf{c}_k\right) &= \frac{1}{Z''_{\tau,\pi}} \sum_{\mathbf{s}_{\tau-1}} p\left(\mathbf{s}_\tau|\mathbf{s}_{\tau-1},\pi\right) m_r^{\tau-1}\left(\mathbf{s}_{\tau-1}|\mathbf{c}_k\right) m_\pi^{\tau-2}\left(\mathbf{s}_{\tau-1}|\mathbf{c}_k\right),
\end{aligned}
\tag{3.17}
$$

where

$$
\ln p'\left(\mathbf{r}_\tau|\mathbf{s}_\tau,\mathbf{c}_k\right) = \int d\phi q\left(\phi\right) \ln p\left(\mathbf{r}_\tau|\mathbf{s}_\tau,\phi,\mathbf{c}_k\right) \tag{3.18}
$$

the free energy mandated that we average out the dependency on ϕ.

The posterior beliefs over policies given some context are calculated as

$$
\begin{aligned}
\ln q\left(\pi|\mathbf{c}_k\right) \propto & \int d\theta q\left(\theta\right) \ln p\left(\pi|\theta,\mathbf{c}_k\right) + \int d\phi q\left(\phi\right) \ln p\left(\mathbf{s}_{1:t},\mathbf{r}_{1:t}|\pi,\phi,\mathbf{c}_k\right) \\
& - \int d\phi q\left(\phi\right) \sum_{\mathbf{s}_{t+1:T},\mathbf{r}_{t+1:T}} q\left(\mathbf{s}_{t+1:T},\mathbf{r}_{t+1:T}|\pi,\mathbf{c}_k\right) \ln \frac{q\left(\mathbf{s}_{t+1:T},\mathbf{r}_{t+1:T}|\pi,\mathbf{c}_k\right)}{p\left(\mathbf{s}_{t+1:T},\mathbf{r}_{t+1:T}|\pi,\phi,\mathbf{c}_k\right)} \\
\propto & \ln p'\left(\pi|\mathbf{c}_k\right) + \sum_{m=1}^{t} \ln p\left(\mathbf{s}_m|\mathbf{s}_{m-1},\pi\right) - \ln Z_\tau^\pi - \sum_{\tau=t+1}^{T} \ln Z''_{\tau,\pi} \\
\propto & \ln p'\left(\pi|\mathbf{c}_k\right) - F\left(\pi|\mathbf{c}_k\right) \\
q\left(\pi|\mathbf{c}_k\right) \propto & p'\left(\pi|\mathbf{c}_k\right) \exp\left(-F\left(\pi|\mathbf{c}_k\right)\right)
\end{aligned}
\tag{3.19}
$$

where $p'\left(\pi|\mathbf{c}_k\right)$ is the marginalized prior over policies, and $F\left(\pi|\mathbf{c}_k\right)$ is the policy-specific free energy in a given context (see (Schwöbel et al., 2018)).

The posterior over the parameters θ of the prior over policies can be derived as

$$
\begin{aligned}
\ln q\left(\theta\right) &\propto \ln p\left(\theta\right) + \sum_{\pi,\mathbf{c}_k} q\left(\pi|\mathbf{c}_k\right) q\left(\mathbf{c}_k\right) \ln p\left(\pi|\theta,\mathbf{c}_k\right) \\
&\propto \ln\left(\frac{1}{B\left(\alpha\right)} \prod_{l,n} \theta_{ln}^{\alpha_{ln}^{k-1}-1}\right) + \sum_{\pi,\mathbf{c}_k} q\left(\pi|\mathbf{c}_k\right) q\left(\mathbf{c}_k\right) \ln\left(\prod_{l,n} \theta_{ln}^{\delta_{l,\pi}\delta_{n,c_k}}\right) \\
&\propto \ln\left(\prod_{l,n} \theta_{ln}^{\alpha_{ln}^{k-1}-1}\right) + \sum_{l,n} q\left(\pi = l|\mathbf{c}_k = n\right) q\left(\mathbf{c}_k = n\right) \ln\left(\theta_{ln}\right) \\
&\propto \ln\left(\prod_{l,n} \theta_{ln}^{\alpha_{ln}^{k-1}-1}\right) + \ln\left(\prod_{l,n} \theta_{ln}^{q\left(\pi=l|\mathbf{c}_k=n\right)q\left(\mathbf{c}_k=n\right)}\right) \\
&\propto \ln\left(\prod_{l,n} \theta_{ln}^{\alpha_{ln}^{k-1}-1+q\left(\pi=l|\mathbf{c}_k=n\right)q\left(\mathbf{c}_k=n\right)}\right) \\
q\left(\theta\right) &= \frac{1}{B\left(\alpha^k\right)} \prod_{l,n} \theta_{ln}^{\alpha_{ln}^{k}-1} \\
\alpha_{ln}^k &= \alpha_{ln}^{k-1} + q\left(\pi = l|\mathbf{c}_k = n\right) q\left(\mathbf{c}_k = n\right)
\end{aligned}
\tag{3.20}
$$

and is itself again a Dirichlet distribution with updated pseudo counts α^k. These are updated by adding the posterior over policies times the posterior over context. At the end of an episode, the pseudo count will be increased by 1 for the policy which has been followed in the context the agent inferred to be in.

The posterior over the parameters ϕ of the outcome rules can be derived as

$$\begin{aligned}
\ln q\left(\phi\right) &\propto \ln p\left(\phi\right) + \sum_{\mathbf{c}_k} q\left(\mathbf{c}_k\right) \ln p\left(\mathbf{r}_{1:t} | \mathbf{s}_{1:t}, \phi, \mathbf{c}_k\right) \\
&\propto \ln p\left(\phi\right) + \sum_{m=1}^{t} \sum_{\mathbf{c}_k} q\left(\mathbf{c}_k\right) \ln p\left(\mathbf{r}_m | \mathbf{s}_m, \phi, \mathbf{c}_k\right) \\
&\propto \ln \left(\frac{1}{B\left(\beta\right)} \prod_{i,j,n} \phi_{ijn}^{\beta_{ijn}^{k-1}-1} \right) + \sum_{m=1}^{t} \sum_{\mathbf{c}_k} q\left(\mathbf{c}_k\right) \ln \left(\prod_{i,j,n} \phi_{ijn}^{\delta_{i,\mathbf{r}_m} \delta_{j,\mathbf{s}_m} \delta_{n,c_k}} \right) \\
&\propto \ln \left(\prod_{i,j,n} \phi_{ijn}^{\beta_{ijn}^{k-1}-1} \right) + \sum_{m=1}^{t} \sum_{i,j,n} q\left(\mathbf{c}_k = n\right) \ln \left(\phi_{ijn}^{\delta_{i,\mathbf{r}_m} \delta_{j,\mathbf{s}_m}} \right) \\
&\propto \ln \left(\prod_{i,j,n} \phi_{ijn}^{\beta_{ijn}^{k-1}-1} \right) + \ln \left(\prod_{i,j,n} \phi_{ijn}^{q(\mathbf{c}_k=n) \sum_m \delta_{i,\mathbf{r}_m} \delta_{j,\mathbf{s}_m}} \right) \\
&\propto \ln \left(\prod_{i,j,n} \phi_{ijn}^{\beta_{ijn}^{k-1}-1+q(\mathbf{c}_k=n) \sum_m \delta_{i,\mathbf{r}_m} \delta_{j,\mathbf{s}_m}} \right) \\
q\left(\phi\right) &= \frac{1}{B\left(\beta^k\right)} \prod_{i,j,n} \phi_{ijn}^{\beta_{ijn}^{k}-1} \\
\beta_{ijn}^{k} &= \beta_{ijn}^{k-1} + q\left(\mathbf{c}_k = n\right) \sum_{m=1}^{t} \delta_{i,\mathbf{r}_m} \delta_{j,\mathbf{s}_m}
\end{aligned} \tag{3.21}$$

Lastly, we want to derive the posterior over contexts

$$\begin{aligned}
\ln q\left(\mathbf{c}_k\right) \propto & \ln p'\left(\mathbf{c}_k\right) - \int d\theta q\left(\theta\right) \sum_{\pi} q\left(\pi | \mathbf{c}_k\right) \ln \frac{q\left(\pi | \mathbf{c}_k\right)}{p\left(\pi | \theta, \mathbf{c}_k\right)} \\
& + \int d\phi q\left(\phi\right) \sum_{\pi} q\left(\pi | \mathbf{c}_k\right) \ln p\left(\mathbf{s}_{1:t}, \mathbf{r}_{1:t} | \pi, \phi, \mathbf{c}_k\right) \\
& - \int d\phi q\left(\phi\right) \sum_{\pi} q\left(\pi | \mathbf{c}_k\right) \sum_{\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T}} q\left(\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T} | \pi, \mathbf{c}_k\right) \ln \frac{q\left(\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T} | \pi, \mathbf{c}_k\right)}{p\left(\mathbf{s}_{t+1:T}, \mathbf{r}_{t+1:T} | \pi, \phi, \mathbf{c}_k\right)} \\
\propto & \ln p'\left(\mathbf{c}_k\right) - D_{KL}\left[q\left(\pi | \mathbf{c}_k\right) | p'\left(\pi | \mathbf{c}_k\right) \right] + \int d\phi q\left(\phi\right) \sum_{m=1}^{t} \ln p\left(\mathbf{r}_m | \mathbf{s}_m, \phi, \mathbf{c}_k\right) - \ln Z_{\tau}^{\pi} - \sum_{\tau=t+1}^{T} \ln Z_{\tau,\pi}'' \\
\propto & \ln p'\left(\mathbf{c}_k\right) - D_{KL}\left[q\left(\pi | \mathbf{c}_k\right) | p'\left(\pi | \mathbf{c}_k\right) \right] - \sum_{\pi} q\left(\pi | \mathbf{c}_k\right) F\left(\pi, \mathbf{c}_k\right) \\
q\left(\mathbf{c}_k\right) \propto & p'\left(\mathbf{c}_k\right) \exp\left(-F\left(\mathbf{c}_k\right)\right)
\end{aligned} \tag{3.22}$$

with context-specific free energy $F\left(\mathbf{c}_k\right)$. Note, that we set

$$p'\left(\mathbf{c}_k\right) = \sum_{\mathbf{c}_{k-1}} q\left(\mathbf{c}_{k-1}\right) p\left(\mathbf{c}_k | \mathbf{c}_{k-1}\right) \tag{3.23}$$

As most of the posteriors described here are interdependent on each other, one has to iterate over their updates until convergence. Practically, we only used one iteration step: We

used the priors over θ, ϕ and $\mathbf{c}_k$ to calculate the posterior over policies. Then we calculated the posteriors over θ and ϕ, which were then used to calculate the posterior over contexts. We evaluated if this procedure is equivalent to a full iteration until convergence and found that the resulting posteriors only differed by less than 1% of their values.

3.8.2 Agent and task setup

The generative process of the habit learning task (Section 3.4.2) was set up as follows:

- An episode has length $T = 2$.
- There are 200 episodes so that $k \in [1, 200]$
- There are $n_s = 3$ states $\mathcal{S} = \{s_1, s_2, s_3\}$, where s_1 is the state where lever 1 distributes a reward, s_2 is the state where lever 2 distributes a reward, and state s_3 is the starting state in front of the two levers.
- There are $n_r = 3$ rewards $\mathcal{R} = \{r_1, r_2, r_3\}$, where r_1 is the reward payed out by lever 1, r_2 is the reward payed out by lever 2, and r_3 is the no-reward.
- There are $n_a = 2$ actions $\mathcal{A} = \{a_1, a_2\}$, where a_1 leads to state s_1, and a_2 leads to state s_2 from any starting state.
- There are $n_c = 2$ contexts $\mathcal{C} = \{c_1, c_2\}$ which amount to lever 1 or lever 2 being the better arm, respectively.
- The state transitions are set up to be deterministic: $\mathcal{T}_s\left(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t = a_1\right) = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ and $\mathcal{T}_s\left(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t = a_1\right) = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$ so that a_1 leads to state s_1 from any starting state, and a_2 to s_2, while s_3 can not be reached.
- The reward generation rules are as depicted in Figure 3.3b. Mathematically, the reward generation in the training phase as $\mathcal{T}_r\left(\mathbf{r}_t|\mathbf{s}_t, \mathbf{c}_k\right) = \begin{pmatrix} \nu & 0 & 0 \\ 0 & 1-\nu & 0 \\ 1-\nu & \nu & 0 \end{pmatrix}$ for $k \in \left[1, d_{training}\right]$, where ν is the probability of lever 1 distributing a reward. In the extinction phase, the reward probabilities switch, so that $\mathcal{T}_r\left(\mathbf{r}_t|\mathbf{s}_t, \mathbf{c}_k\right) = \begin{pmatrix} 1-\nu & 0 & 0 \\ 0 & \nu & 1 \\ \nu & 1-\nu & 0 \end{pmatrix}$ for $k \in \left[d_{training} + 1, d_{training} + 100\right]$
- The context transitions are deterministic and happen after the training, so that $\mathcal{T}_c\left(\mathbf{c}_{k+1}|\mathbf{c}_k\right) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ for $k \in \left\{1, \ldots, d_{\text{training}}, d_{\text{training}} + 2, \ldots, d_{training} + 100\right\}$ and $\mathcal{T}_c\left(\mathbf{c}_{k+1}|\mathbf{c}_k\right) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ for $k = d_{\text{training}} + 1$.

In each episode k, the agent starts at $t = 1$ in the state s_3 in front of the levers.

The agent's generative model is set up to reflect the generative process, or learn the respective quantities:

- The agent knows it starts in state s_3 in each episode, so we set the prior of the starting state as $p\left(\mathbf{s}_1|\mathbf{s}_0,\pi\right) = p\left(\mathbf{s}_1\right) = (0,0,1)^T$
- As we set $T = 2$, policies and actions map one to one, so that $len\left(\pi\right) = 1$ and $n_\pi = 2$. This means, $\pi_1 = a_1$ and $\pi_2 = a_2$
- We assume the agent knows the state transitions instead of learning those, so that $p\left(\mathbf{s}_{t+1}|\mathbf{s}_t,\pi\right) = \mathcal{T}_s\left(\mathbf{s}_{t+1}|\mathbf{s}_t,\mathbf{a}_t\right)$
- The pseudo counts β_{ijn}^k which are used to parameterize the outcome rules for reward i and state j in context n, are initialized as $\beta_{ijn}^0 = 1$ for all i,j,n
- The pseudo counts α_{ln}^k which parameterize the prior over actions for policy l in context n are initialized as $\alpha_{ln}^0 = \alpha_{\text{init}} = \frac{1}{h}$ using the habitual tendency h and are initialized the same for all l,n.
- We set the agent's representation of context transitions, i.e. temporal stability as $p\left(\mathbf{c}_{k+1}|\mathbf{c}_k\right) = \begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix}$ for $k = d_{\text{training}} + 1$. Here, the agent assumes that both contexts are equally stable and change once in 100 trials.
- Finally, we set the agents preference for outcomes as $p\left(R = 1|\mathbf{r}_\tau\right) = (0.495, 0.495, 0.01)^T$, so that the agent prefers the rewards of levers 1 and 2 equally, but dislikes the no-reward r_3. In the contingency degradation tasks, these values are kept constant. In the outcome devaluation task (Section 3.4.7), the preference for outcomes was reset in the extinction phase as $p\left(R = 1|\mathbf{r}_\tau\right) = (0.01942, 0..96117, 0.01942)^T$, which effectively devalues the reward for lever 1 and keeps the ratio of desirability between reward and no reward unchanged.

4 Discussion

4.1 Summary

In this thesis, I derived computational hypotheses and definitions of habit learning from the agreed upon operational definitions used in experiments. I combined three key features to explain habit learning: that it is context dependent and of hierarchical nature, that habits can be expressed as chunked action sequences, and that they are learned through mere repetition, analogous to motor skill learning. Furthermore I proposed that the balancing can be done by using Bayes' rule, where uncertainty-based weighting of control contributions arises from simple multiplication. Should this model hold under experimental scrutiny, the novel mechanistic understanding would provide profound implications.

In Chapter 2, I proposed an improved inference scheme for sequential inference in the active inference framework to solve a Markov decision process. Here, the second order Bethe approximation was used to infer states and observations in a partially observable Markov decision process. This allowed for the use the belief propagation algorithm to calculate beliefs at the minimum of the variational free energy. Using a grid world as a toy setup for a simulated agent, I could show that this improved sequential inference leads to an increase in goal-reaching performance in environments with state transition uncertainty as well as observation uncertainty, when compared to the mean-field approximation which had been typically used before.

In Chapter 3, I proposed a hierarchical Bayesian model of habit learning. On the lower level of the hierarchy, habits are interpreted as a prior over policies, i.e. sequences of actions, which are learned based on repetition of previous behavior. Goal-directed control arises from an explicit Bayesian evaluation of a Markov decision process in the likelihood, where the algorithm from Chapter 2 was used. Both control signals are automatically weighted based on their respective uncertainties through a simple multiplication in the posterior, from which actions are sampled and executed. Interestingly, the prior constrains which policies will be evaluated in the likelihood, so that habit learning also constrains the decision tree and the computational cost of the goal-directed system. Additionally, the model contains a context in the higher level of the hierarchy which determines action-outcome contingencies. As a consequence, an agent may know that it can be in different contexts and learn action-outcome contingencies as well as habits for specific contexts. The model has a free parameter, the habitual tendency h, which can capture inter-individual differences in habit learning and resulting habit strength.

Using this model for simulated agents, I showed that this habit learning mechanism replicates the key characteristics of habit learning as discussed in Chapter 1 (see also below), and how different habit strengths emerge from different habitual tendencies and experimental manipulations.

4.2 Contributions

As outlined in the Introduction, the habit learning model proposed in this thesis is the first one to capture all of the following key habit learning characteristics: (i) habitual behavior is insensitive to outcome devaluation and contingency degradation, (ii) habit strength increased with training duration, (iii) habits are more resource efficient than goal-directed control, (iv) habit strength increases with decreased action-outcome contingency, and (v) habits are context-sensitive and can be quickly recalled in a familiar environment. Additionally, the model offers a simple way to balance control contributions.

Out of this list of features, (i) and (ii) are the most characteristic for habits, as they correspond to the operational definitions of habits from the experimental literature. Hence, all habit learning models originally set out to describe and model these effects. For the well cited model-free/model-based habit learning model it has been shown however, that model-free contributions to behavior do not predict insensitivity to outcome devaluation in humans, which make it unlikely that the model-free interpretation of habit learning holds. This argument holds for the plan-until-until habit model as well, which also relies on model-free learning, where stronger habit contributions were compared to a decreased planning depth. Unfortunately, the authors did not investigate how this changes with amount of training, so it is unclear whether this model describes effect (ii).

Many models have investigated why habits may be inflexible, but have failed to provide a convincing account of (iii) how habits are more resource efficient than reliance on goal-directed control. When proposing a mechanistic account of habits, this is a very important aspect, as the resource and time efficiency of habits is the very reason why an agent should switch away from more accurate goal-directed control to less accurate habitual control. The model-free/model-based model (Daw et al., 2005, 2011), the value-free/value-based (Miller et al., 2019), as well as the Bayesian proposals based on model comparison (FitzGerald et al., 2014a; K. Friston, FitzGerald, Rigoli, Schwartenbeck, O'Doherty, & Pezzulo, 2016) all need to evaluate the full goal-directed Markov decision process in order to obtain goal-directed control contributions, even if they are low. This defies the purpose of habit learning and therefore makes it improbable that these model are appropriate mechanistic explanations. The hierarchical model (Dezfouli & Balleine, 2012, 2013), the plan-until-habit model (Keramati et al., 2016), and the Bayesian caching model (Maisto et al., 2019) provide accounts of how habitual control improves reaction time and mental resource consumption. In the model proposed in this thesis, weighting is done in the Bayesian way: by multiplication of the prior and the posterior. This offers a simple and elegant way to achieve uncertainty-based control contributions. An interesting consequence of this method is, that policies which have a prior of zero, will always have a posterior of zero and will never be chosen. Consequently, the prior, i.e. the habit, provides a way for an agent to learn which policies can be excluded from goal-directed evaluation a priori. So the habit not only offers a quick way to evaluate actions based on past experience, but also guides and constrains the goal-directed evaluation in the likelihood. As a consequence, the habit model does fulfill its purpose of yielding a quicker and more resource efficient way of action evaluation.

The action-outcome contingency (iv) is defined as the spatio-temporal correlation of actions and their consequence. It is well known, that the strength of conditioning increases with the action-outcome contingency (Mazur, 2015; Gluck et al., 2016). For habits, which are measured in extinction after the conditioning or training phase, the opposite holds: habit strength is increased with decreased action-outcome contingency, in which case conditioning as well as extinction of behavior takes longer. Experimentally, this has been mainly shown through the use of different reinforcement schedules in animals, where schedules that have a lower contingency, such as variable interval schedules, achieve higher habitization (Yin & Knowlton, 2006). This is an effect which is hard to reproduce with any habit learning model based on reward history, as for example model-free/model-based model. Here, the decreased contingency would decrease model-free as well as model-based control contributions. Mechanistic descriptions based on repetition learning, as the value-free/value-based model and the one proposed in this thesis, can account better for this effect, as the decreased contingency decreases goal-directed control contributions, but not habitual contributions.

Lastly, the context sensitivity (v) is a habit characteristic which has been a matter for debate with regard to modeling. None of the previous models except for the context models (Redish et al., 2007; Gershman et al., 2010) have explicitly introduced and dealt with contexts. As a result, habit learning models based on model-free temporal difference reinforcement learning can only describe behavior in the extinction phase where behavior is slowly unlearned, but not the quick recall of behavior upon entering a known context (Redish et al., 2007). As a remedy, contexts have been proposed to be interpreted in terms of the states of the Markov decision process (Dolan & Dayan, 2013). In these approaches, a flat model is used, where the states encode the context, and different states lead to different outcomes, depending on the action. Conversely, the model proposed here explicitly includes context as a separate variable at the higher level of the hierarchical model, more akin to the context models, explicitly encoding contexts and determining the action-outcome contingencies in each state. Concretely, this equips an agent with the ability to learn, encode, and recall different environments as different Markov decision processes and flexibly store and load this encoding upon leaving and entering a familiar environment. Given the hierarchical nature of the brain, and the multitude of contexts and environments each living agent encounters in their everyday lives, this may be a more convincing mechanistic description of context sensitivity.

Taken together, I was able to show that a quantitative description of habit learning based on the ideas that habits are hierarchically organized sequences of actions in a Bayesian model and learned based on repetition, can replicate these five key characteristics of habit learning. Unfortunately it is currently unclear if these hypotheses, which were based on animal research, generalize to human habit learning. Nonetheless, given no satisfactory translation of animal habit learning experiments to human experimental paradigms has yet been achieved (de Wit et al., 2018; Friedel et al., 2014; Gillan et al., 2015), the proposed mechanistic description may guide more successful human task development, where successful habit induction in humans could be achieved through the components hypothesised in this model. Concretely, instead of using single actions with high action-outcome contingencies as in outcome devaluation tasks, future experiments could use sequences of actions with probabilistic outcomes to induce habits. In contrast to the two-step task, habit learning could be facilitated by designing a task such that the same action sequence reliably has the highest success rate in a specific context, which could lead to behavioral automatization through extended training. Habits could be probed by testing whether participants continue to execute this sequence even after a contingency change in an extinction phase.

4.3 Implications

Besides the potential for aiding task development, the model may help to measure habits in a dimensional way. The current habit strength corresponds to the precision of the prior over policies, where higher precision leads to larger control control contributions, which may be compared to the precision of the goal-directed control contributions to a measure of relative habit strength. When fitting the model to data, this may yield the opportunity to independently get a measure for both control contributions as well as their relative strength. As a result, using this model in future experiments to asses habit strength may make it possible to not only measure if a participant relies on habits or not, as measured by failure to adapt to outcome devaluation for example, but also to assess how much each individual relies on habits at any moment in time, as a sort of dynamical measure of habit strength. So far, experimental probes of habits had to rely on over training animals or human participants to induce habits. This novel analysis technique could open the possibility to study moderate training periods, which on one hand is easier to achieve experimentally due to taking less time, and on the other hand may lead to larger individual differences in habit control strengths compared to the over-trained case. As a result, probing for increased or decreased reliance on habits in groups with mental disorders would be facilitated, see Discussion of Chapter 3.

While I do propose a computational and mechanistic explanation of habit learning, it is not shown in this thesis how this description should map to neuronal architecture. One important aspect of achieving a plausible mapping to neuronal architecture is how the proposed Bayesian computations can be implemented into neural networks. For the sequential inference based on the belief propagation algorithm, theoretical studies were able to show concrete implementations of belief propagation in rate coded (Shon & Rao, 2005; Ott & Stoop, 2007) as well as spiking neural networks (Deneve, 2005), making it plausible that this may be implemented in the brain. For the more complex hierarchical habit learning model the picture is not as clear. On one hand, there are neural network implementations based on recurrent neural networks, which were able to learn action-outcome contingencies in a context-dependent manner (Butz et al., 2019). The authors showed in simulations that an agent can learn how to behave in different contexts and adapt its behavior upon encountering a new context. While based on a similar graph, it is unclear if this neural network is a concrete instance of the Bayesian processing described in this work which needs to be investigated in the future. Additionally, the proposed model contains the habitual prior over policies, but it is uncertain what a prior means in terms of neural networks. Some studies propose that the spontaneous activity in a spiking neural network may correspond to a Bayesian prior (Berkes, Orbán, Lengyel, & Fiser, 2011), where the inputs to the network correspond to the likelihood, and the outputs to the posterior. This is a plausible interpretation of a prior, but taking the arguments provided above together, there is currently no concrete translation of the Bayesian habit learning model into a neural network. Nonetheless, the components discussed above could be taken together to build a neural network translation of the Bayesian model proposed in this thesis.

Another important aspect of achieving a plausible mapping to neuronal architecture is whether the mechanistic account fits to the neurobiological underpinnings of habit learning, as outlined in Section 1.2. The habit learning model contains various aspects such as learning, storing, and loading of contextual action-outcome contingencies and habits, and how a habit may guide and constrain the goal-directed evaluation. These aspects were also identified in several neuronal findings regarding habit learning: the dorsomedial striatum is implicated in

learning and evaluating action-outcome contingencies (Yin et al., 2005; E. M. Tricomi et al., 2004), while the orbitofrontal cortex is thought to guide context-dependent retrieval of the contingencies (Gremel & Costa, 2013; Parkes et al., 2018), and the medial prefrontal cortex has been found to store and encode valuation signals (Daw et al., 2006; B. W. Balleine & O'Doherty, 2010). The dorsolateral striatum is heavily implicated in habit learning (Reep et al., 2003; Yin et al., 2004), where a task bracketing activity has been found and was shown to correlate with behavioral automaticity (Smith & Graybiel, 2013, 2014). These findings may correspond to the way the habitual prior guides loading and evaluation of action-outcome contingencies at the start of a behavioral episode. For a more detailed description of the potential mapping to the neurobiological findings on habit learning, see the Discussion in Chapter 3.

A better mechanistic description of habit learning, as well as an improved understanding of the corresponding neurobiology will also have strong implications for better understanding and treating mental disorders which are thought to be correlated with a maladaptive balance of habitual and goal-directed control, such as addition and OCD. In the habit learning model, the habitual tendency parameter h offers a way to describe interindividual differences in habit learning trajectories: An individual with a high habitual tendency will exhibit stronger habitual control with less repetition of the same behavior compared to an individual with a low habitual tendency. Additionally, the model is able to describe how different habit learning trajectories emerge as a consequence of the individual habit tendency h, properties of the environment, like for example its stability, as well as their interaction. In addiction for example, addicted individuals are thought to shift from goal-directed to habitual control as they form the addiction (Volkow & Morales, 2015; Everitt & Robbins, 2005, 2016), but these trajectories may differ between individuals. Therefore, the model may offer a novel way to better understand these individual trajectories, for a more detailed discussion see the Discussion section in Chapter 3.

4.4 Interpretation

Apart form the concrete interpretations of the model in the scope of instrumental and habit learning, the proposed model can be interpreted in the scope of broader psychological concepts such as motivation and executive function or cognitive control. Studies in rodents investigated the influence of motivational states on instrumental learning and habits (B. Balleine, 1992; Lopez, Balleine, & Dickinson, 1992; Dickinson, Balleine, Watt, Gonzalez, & Boakes, 1995; Dickinson & Balleine, 1994; B. W. Balleine & O'Doherty, 2010) and found that motivational changes influence habit learning trajectories. For moderate training durations, a shift from increased motivation during the conditioning phase, induced by hunger or thirst, to decreased motivation in extinction, in a non-deprived state, decreases the repetition of conditioned behavior in extinction. For extended training however, this effect vanishes. Following the proposition by Niv, Joel, and Dayan (2006), in a computational model the motivational state may be mapped to the utility of an outcome, which corresponds to the prior preference over outcomes in Bayesian and active inference models. Such a shift in the utility would lead to an increased valuation of an action in the goal-directed evaluation and thus yield a more pronounced likelihood, while the habitual prior over actions would only change if an action has been performed more or less frequently. As a result, the goal-directed likelihood receives an increased weighting the posterior over actions, leading to an increased weight in action selection, as long as it is not solely dominated by the prior, as is the case for moderate training

durations. For extended training durations, the prior becomes so pronounced and dominant that changes in the goal-directed evaluation have little influence in the balancing of control, which renders behavior in extinction almost insensitive to motivational changes. This fits very well to the experimental findings on motivation in the animal literature.

Integrating the model into broader concepts such as executive function and cognitive control (Goschke, 2014; "The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis", 2000), arguments can be made the model captures effects such as inhibition and updating, two prominent features of higher order cognitive function. According to the model, the prior over policies is loaded upon receiving either a context cue or inferring the current context. While not concretely built into the model, it is imaginable that this may happen in the cortico-basal ganglia-loop centered on the dorsolateral striatum. Here, the loading could be done rather quickly, so that a fast motor response can be prepared, akin to a reflex or impulse. At the same time, the current action-outcome contingencies could be loaded into the goal-directed loop centered on the dorsomedial striatum. The prior could then determine which policies are to be evaluated in a goal-directed manner, so that policies which have a high a priori probability of being executed are evaluated first. Only after some time, when sufficiently many policies have been evaluated, would the likelihood be fully known and be able to override the impulse of following the habit, leading to suppression or inhibition of the habitual response. In this interpretation, the initial loading of the prior would correspond to an impulse of following the habit, while the slower goal-directed evaluation which may override the impulse would correspond to inhibition. This could also explain why habits are more prevalent under time pressure: If there is not enough time to evaluate the full likelihood, an agent would need to rely on the faster, pre-loaded prior. Furthermore, the (re-)loading of the contextual habit and action-outcome contingencies upon receiving a contextual cue or inferring the context would correspond to updating of the neural representation when it is warranted.

4.5 Limitations

While the model proposed seems able to explain many key properties of habit learning, it comes with limitations that future research needs to resolve. One potential limitation is that it is time discrete, i.e. it rests on the assumption that actions are evaluated and chosen in discrete time steps. In many human experiments, this would not be an issue as they are often based on trials. In the animal experiments as discussed in Section 1.1, habit induction critically depends on the reinforcement schedule used, which can be based on intervals and exact time points at which the reward is delivered, and behavior is measured in the form of response rates. Both are aspects that would require a time-continuous model which is able to model the contingency not only based on conditional probabilities, but the exact timing of the reward distribution and lever press could be included. Consequently, the model can only indirectly capture the properties of such reinforcement schedules via the conditional action-outcome probabilities, but it can not directly explain why an animal would choose an increased lever press rate. Miller et al. (2019) tried to tackle this limitation in their time-discrete model by not only letting an agent choose which action to take, but also at which rate it wants to apply this action. A similar approach could be used to resolve this issue in the proposed model.

Another consequence of the discrete time is that the model so far does not offer a way to describe reaction times. This is important, as the decrease of reaction time is a known property of increased behavioral automaticity (Seger & Spiering, 2011). Furthermore, many studies

have shown that reaction times themselves can be indicative of the underlying cognitive process, like e.g. in perceptual decision making (Heekeren, Marrett, & Ungerleider, 2008). There are several interesting approaches one could choose to obtain a reaction time from the posterior over policies: It may be that the reaction time increases with the entropy of the posterior, i.e. reaction times may be higher the more uncertain the evaluation is. To potentially achieve a similar effect, one may propose a specific sampling algorithm how the agent samples actions from the posterior. A good candidate would be Hamilton Monte-Carlo sampling (R. M. Neal, 2011), where random variables are treated as state space variables of a dynamical system which is simulated according to Hamiltonian dynamics to achieve the sampling. Interestingly, it has been shown that neural networks could potentially implement this method (Aitchison & Lengyel, 2016), making this a viable approach. Another additional factor which surely would influence reaction times is the model complexity determining the time an agent would need to plan through the Markov decision process. Here, reaction times should increase with the number of potential policies, the number of states, and the planning depth.

Future work may investigate how the model can be made time continuous and incorporate reaction times. Additionally, future experimental studies will reveal whether this model fits habit learning behavior in animals and humans better than the previous approaches. Despite these limitations, the model proposed in this thesis could prove useful to better understand habit learning in a mechanistical sense, and may help to aide the development of improved behavioral paradigms to induce habits in humans. A better understanding of habits may have implications for mental disorders that are thought to be accompanied by a maladapted balance between habitual and goal-directed control.

References

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B*, *34*(2b), 77–98.

Aitchison, L., & Lengyel, M. (2016). The hamiltonian brain: efficient probabilistic inference with excitatory-inhibitory neural circuit dynamics. *PLoS computational biology*, *12*(12).

Akam, T., Costa, R., & Dayan, P. (2015). Simple plans or sophisticated habits? state, transition and learning interactions in the two-step task. *PLoS computational biology*, *11*(12).

Alloway, K. D., Smith, J. B., Mowery, T. M., & Watson, G. D. (2017). Sensory processing in the dorsolateral striatum: the contribution of thalamostriatal pathways. *Frontiers in systems neuroscience*, *11*, 53.

Arthur, W. B. (1994). Inductive reasoning and bounded rationality. *The American economic review*, *84*(2), 406–411.

Astrom, K. J. (1965). Optimal control of markov decision processes with incomplete state estimation. *Journal of mathematical analysis and applications*, *10*(1), 174–205.

Attias, H. (2003a). Planning by probabilistic inference. In *Aistats.*

Attias, H. (2003b). Planning by Probabilistic Inference. In *Proc. of the 9th int. workshop on artificial intelligence and statistics.*

Baker, C., Saxe, R., & Tenenbaum, J. B. (2006). Bayesian models of human action understanding. In *Advances in neural information processing systems* (pp. 99–106).

Balleine, B. (1992). Instrumental performance following a shift in primary motivation depends on incentive learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *18*(3), 236.

Balleine, B. W. (2019). The meaning of behavior: discriminating reflex and volition in the brain. *Neuron*, *104*(1), 47–62.

Balleine, B. W., & Killcross, S. (2006). Parallel incentive processing: an integrated view of amygdala function. *Trends in neurosciences*, *29*(5), 272–279.

Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, *35*(1), 48–69.

Barnes, T. D., Kubota, Y., Hu, D., Jin, D. Z., & Graybiel, A. M. (2005). Activity of striatal neurons reflects dynamic encoding and recoding of procedural memories. *Nature*, *437*(7062), 1158.

Barto, A. G., Sutton, R. S., & Watkins, C. (1989). *Learning and sequential decision making*. University of Massachusetts Amherst, MA.

Beal, M. J. (2003). *Variational algorithms for approximate bayesian inference*. University of London London.

Behrens, T. E., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. (2008). Associative learning of social value. *Nature*, *456*(7219), 245.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature neuroscience*, *10*(9), 1214.

Berkes, P., Orbán, G., Lengyel, M., & Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, *331*(6013), 83–87.

Berridge, K. C., & Kringelbach, M. L. (2015). Pleasure systems in the brain. *Neuron*, *86*(3), 646–664.

Bethe, H. (1931). Zur theorie der metalle. *Zeitschrift für Physik A Hadrons and Nuclei*, *71*(3), 205–226.

Bethe, H. A. (1935). Statistical theory of superlattices. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *150*(871), 552–575.

Bishop, C. M. (2006a). *Pattern recognition and machine learning*. springer.

Bishop, C. M. (2006b). *Pattern recognition and machine learning*. springer.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859-877. Retrieved from `http://dx.doi.org/10.1080/01621459.2017.1285773` doi: 10.1080/01621459.2017.1285773

Botvinick, M., & Toussaint, M. (2012a). Planning as inference. *Trends in cognitive sciences*, *16*(10), 485–488.

Botvinick, M., & Toussaint, M. (2012b, oct). Planning as inference. *Trends in Cognitive Sciences*, *16*(10), 485–488. Retrieved from `https://doi.org/10.1016%2Fj.tics.2012.08.006` doi: 10.1016/j.tics.2012.08.006

Bouton, M. E. (2019). Extinction of instrumental (operant) learning: interference, varieties of context, and mechanisms of contextual control. *Psychopharmacology*, *236*(1), 7–19.

Bouton, M. E., & Bolles, R. C. (1979). Contextual control of the extinction of conditioned fear. *Learning and motivation*, *10*(4), 445–466.

Brighton, H., & Gigerenzer, G. (2008). Bayesian brains and cognitive mechanisms: Harmony or dissonance. *The probabilistic mind: Prospects for Bayesian cognitive science, ed. N. Chater & M. Oaksford*, 189–208.

Butz, M. V. (2016). Toward a unified sub-symbolic computational theory of cognition. *Frontiers in psychology*, *7*, 925.

Butz, M. V., Bilkey, D., Humaidan, D., Knott, A., & Otte, S. (2019). Learning, planning, and control in a monolithic neural event inference architecture. *Neural Networks*, *117*, 135–144.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181–204.

Colwill, R. M., & Rescorla, R. A. (1988). The role of response-reinforcer associations increases throughout extended instrumental training. *Animal Learning & Behavior*, *16*(1), 105–111.

Cooper, R. P., & Peebles, D. (2015). Beyond single-level accounts: The role of cognitive architectures in cognitive scientific explanation. *Topics in Cognitive Science*, *7*(2), 243–258.

Corbit, L. H. (2018). Understanding the balance between goal-directed and habitual behavioral control. *Current opinion in behavioral sciences*, *20*, 161–168.

Corbit, L. H., & Balleine, B. W. (2015). Learning and motivational processes contributing to pavlovian–instrumental transfer and their neural bases: dopamine and beyond. In *Behavioral neuroscience of motivation* (pp. 259–289). Springer.

Corbit, L. H., Nie, H., & Janak, P. H. (2012). Habitual alcohol seeking: time course and the contribution of subregions of the dorsal striatum. *Biological psychiatry*, *72*(5), 389–395.

Coughlan, J. M., & Ferreira, S. J. (2002). Finding deformable shapes using loopy belief propagation. In *European conference on computer vision* (pp. 453–468).

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., & Friston, K. (2020). Active inference on discrete state-spaces: a synthesis. *arXiv preprint arXiv:2001.07203*.

Danner, U. N., Aarts, H., & de Vries, N. K. (2008). Habit vs. intention in the prediction of future behaviour: The role of frequency, context stability and mental accessibility of past behaviour. *British Journal of Social Psychology*, *47*(2), 245–265.

Daunizeau, J., Den Ouden, H. E., Pessiglione, M., Kiebel, S. J., Stephan, K. E., & Friston, K. J. (2010). Observing the observer (i): meta-bayesian models of learning and decision-making. *PLoS One*, *5*(12), e15554.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, *8*(12), 1704.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

Dayan, P. (2009). Goal-directed control and its antipodes. *Neural Networks*, *22*(3), 213–219.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The helmholtz machine. *Neural computation*, *7*(5), 889–904.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Current opinion in neurobiology*, *18*(2), 185–196.

Deneve, S. (2005). Bayesian inference in spiking neurons. In *Advances in neural information processing systems* (pp. 353–360).

DeRusso, A., Fan, D., Gupta, J., Shelest, O., Costa, R. M., & Yin, H. H. (2010). Instrumental uncertainty as a determinant of behavior under interval schedules of reinforcement. *Frontiers in integrative neuroscience*, *4*, 17.

Deserno, L., Huys, Q. J., Boehme, R., Buchert, R., Heinze, H.-J., Grace, A. A., … Schlagenhauf, F. (2015). Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proceedings of the National Academy of Sciences*, *112*(5), 1595–1600.

Desrochers, T. M., Amemori, K.-i., & Graybiel, A. M. (2015). Habit learning by naive macaques is marked by response sharpening of striatal neurons representing the cost and outcome of acquired action sequences. *Neuron*, *87*(4), 853–868.

de Wit, S., Kindt, M., Knot, S. L., Verhoeven, A. A., Robbins, T. W., Gasull-Camos, J., … Gillan, C. M. (2018). Shifting the balance between goals and habits: Five failures in experimental habit induction. *Journal of Experimental Psychology: General*, *147*(7), 1043.

Dezfouli, A., & Balleine, B. W. (2012). Habits, action sequences and reinforcement learning. *European Journal of Neuroscience*, *35*(7), 1036–1051.

Dezfouli, A., & Balleine, B. W. (2013). Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. *PLoS computational biology*, *9*(12), e1003364.

Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *308*(1135), 67–78.

Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, *22*(1), 1–18.

Dickinson, A., Balleine, B., Watt, A., Gonzalez, F., & Boakes, R. A. (1995). Motivational control after extended instrumental training. *Animal Learning & Behavior*, *23*(2), 197–206.

Dickinson, A., Nicholas, D., & Adams, C. D. (1983). The effect of the instrumental training contingency on susceptibility to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology*, *35*(1), 35–51.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325.

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current opinion in neurobiology*, *22*(6), 1075–1081.

Doshi-Velez, F., Wingate, D., Roy, N., & Tenenbaum, J. B. (2010). Nonparametric bayesian policy priors for reinforcement learning. In *Advances in neural information processing systems* (pp. 532–540).

Doya, K. (2008). Modulators of decision making. *Nature neuroscience*, *11*(4), 410.

Doya, K., Ishii, S., Pouget, A., & Rao, R. P. (2007). *Bayesian brain: Probabilistic approaches to neural coding*. MIT press.

Drake, A. W. (1962). *Observation of a markov process through a noisy channel* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Eppinger, B., Walter, M., Heekeren, H. R., & Li, S.-C. (2013). Of goals and habits: age-related and individual differences in goal-directed decision-making. *Frontiers in neuroscience*, *7*, 253.

Ersche, K., Gillan, C., Jones, S., Williams, G., Ward, L., Luijten, M., … Robbins, T. (2016). Carrots and sticks fail to change behavior in cocaine addiction. *Science*, *352*(6292), 1468–1471.

Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nature neuroscience*, *8*(11), 1481.

Everitt, B. J., & Robbins, T. W. (2013). From the ventral to the dorsal striatum: devolving views of their roles in drug addiction. *Neuroscience & Biobehavioral Reviews*, *37*(9), 1946–1954.

Everitt, B. J., & Robbins, T. W. (2016). Drug addiction: updating actions to habits to compulsions ten years on. *Annual review of psychology*, *67*, 23–50.

Fan, J. L. (2001). Forward-backward algorithm. *Constrained Coding and Soft Iterative Decoding*, 97–116.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2006). Efficient belief propagation for early vision. *International journal of computer vision*, *70*(1), 41–54.

Fino, E., Vandecasteele, M., Perez, S., Saudou, F., & Venance, L. (2018). Region-specific and state-dependent action of striatal gabaergic interneurons. *Nature communications*, *9*(1), 1–17.

FitzGerald, T. H., Dolan, R. J., & Friston, K. J. (2014a). Model averaging, optimal inference, and habit formation. *Frontiers in human neuroscience*, *8*, 457.

FitzGerald, T. H., Dolan, R. J., & Friston, K. J. (2014b). Model averaging, optimal inference, and habit formation. *Frontiers in human neuroscience*, *8*.

FitzGerald, T. H., Hämmerer, D., Friston, K. J., Li, S.-C., & Dolan, R. J. (2017). Sequential inference as a mode of cognition and its correlates in fronto-parietal and hippocampal brain regions. *PLoS computational biology*, *13*(5), e1005418.

Friedel, E., Koch, S. P., Wendt, J., Heinz, A., Deserno, L., & Schlagenhauf, F. (2014). Devaluation and sequential decisions: linking goal-directed and model-based behavior. *Frontiers in human neuroscience*, *8*, 587.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in cognitive sciences*, *13*(7), 293–301.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, *11*(2), 127–138.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience & Biobehavioral Reviews*, *68*, 862–879.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2016). Active inference: A process theory. *Neural computation*.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1521), 1211–1221.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive neuroscience*, *6*(4), 187–214.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2013). The anatomy of choice: active inference and agency. *Frontiers in human neuroscience*, *7*.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2014). The anatomy of choice: dopamine and decision-making. *Phil. Trans. R. Soc. B*, *369*(1655),

20130481.
Friston, K. J., Daunizeau, J., Kilner, J., & Kiebel, S. J. (2010). Action and behavior: a free-energy formulation. *Biological cybernetics*, *102*(3), 227–260.
Friston, K. J., Parr, T., & de Vries, B. (2017). The graphical brain: belief propagation and active inference. *Network Neuroscience*.
Friston, K. J., Rosch, R., Parr, T., Price, C., & Bowman, H. (2017). Deep temporal models and active inference. *Neuroscience & Biobehavioral Reviews*.
Garbusow, M., Schad, D. J., Sommer, C., Jünger, E., Sebold, M., Friedel, E., … Rapp, M. A. (2014). Pavlovian-to-instrumental transfer in alcohol dependence: a pilot study. *Neuropsychobiology*, *70*(2), 111–121.
Gelb, A. (1974). *Applied optimal estimation*. MIT press.
George, D., & Hawkins, J. (2009). Towards a mathematical theory of cortical micro-circuits. *PLoS computational biology*, *5*(10), e1000532.
Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological review*, *117*(1), 197.
Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*, *143*(1), 182.
Gillan, C. M., Otto, A. R., Phelps, E. A., & Daw, N. D. (2015). Model-based learning protects against forming habits. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(3), 523–536.
Gillan, C. M., Papmeyer, M., Morein-Zamir, S., Sahakian, B. J., Fineberg, N. A., Robbins, T. W., & de Wit, S. (2011). Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder. *American Journal of Psychiatry*, *168*(7), 718–726.
Gluck, M. A., Mercado, E., & Myers, C. E. (2016). *Learning and memory*. Worth Publishers.
Gomez-Marin, A., Paton, J. J., Kampff, A. R., Costa, R. M., & Mainen, Z. F. (2014). Big behavioral data: psychology, ethology and the foundations of neuroscience. *Nature neuroscience*, *17*(11), 1455.
Goschke, T. (2014). Dysfunctions of decision-making and cognitive control as transdiagnostic mechanisms of mental disorders: advances, gaps, and needs in current research. *International journal of methods in psychiatric research*, *23*(S1), 41–57.
Graybiel, A. M. (2008). Habits, rituals, and the evaluative brain. *Annu. Rev. Neurosci.*, *31*, 359–387.
Graybiel, A. M., & Grafton, S. T. (2015). The striatum: where skills and habits meet. *Cold Spring Harbor perspectives in biology*, *7*(8), a021691.
Gremel, C. M., & Costa, R. M. (2013). Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. *Nature communications*, *4*, 2264.
Haggard, P. (2019). The neurocognitive bases of human volition. *Annual review of psychology*, *70*, 9–28.
Heekeren, H. R., Marrett, S., & Ungerleider, L. G. (2008). The neural systems that mediate human perceptual decision making. *Nature reviews neuroscience*, *9*(6), 467–479.
Heinz, A., Beck, A., Halil, M. G., Pilhatsch, M., Smolka, M. N., & Liu, S. (2019). Addiction as learned behavior patterns. *Journal of clinical medicine*, *8*(8), 1086.
Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (tec): A framework for perception and action planning. *Behavioral and brain sciences*, *24*(5), 849–878.
Hua, G., Yang, M.-H., & Wu, Y. (2005). Learning to estimate human pose with data driven belief propagation. In *Computer vision and pattern recognition, 2005. cvpr 2005. ieee computer*

society conference on (Vol. 2, pp. 747–754).
Jardri, R., & Denève, S. (2013). Circular inferences in schizophrenia. *Brain*, *136*(11), 3227–3241.
Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, *82*(1), 35–45.
Kaplan, R., & Friston, K. (2017). Planning and navigation as active inference. *bioRxiv*. Retrieved from `https://www.biorxiv.org/content/early/2017/12/07/230599` doi: 10.1101/230599
Keeney, R. L., & Raiffa, H. (1993). *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge university press.
Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS computational biology*, *7*(5), e1002055.
Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proceedings of the National Academy of Sciences*, *113*(45), 12868–12873.
Knill, D. C., & Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, *27*(12), 712–719.
Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244–247.
Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. *Neuron*, *93*(3), 480–490.
Kwisthout, J., & van Rooij, I. (2013). Predictive coding and the bayesian brain: Intractability hurdles that are yet to be overcome. In *Cogsci*.
Kwisthout, J., Wareham, T., & van Rooij, I. (2011). Bayesian intractability is not an ailment that approximation can cure. *Cognitive Science*, *35*(5), 779–784.
Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.
Lally, P., Van Jaarsveld, C. H., Potts, H. W., & Wardle, J. (2010). How are habits formed: Modelling habit formation in the real world. *European journal of social psychology*, *40*(6), 998–1009.
Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, *81*(3), 687–699.
Lee, T. S., & Mumford, D. (2003a). Hierarchical bayesian inference in the visual cortex. *JOSA A*, *20*(7), 1434–1448.
Lee, T. S., & Mumford, D. (2003b). Hierarchical bayesian inference in the visual cortex. *JOSA A*, *20*(7), 1434–1448.
Lim, T., Cardinal, R., Savulich, G., Jones, P., Moustafa, A., Robbins, T., & Ersche, K. (2019). Impairments in reinforcement learning do not explain enhanced habit formation in cocaine use disorder. *Psychopharmacology*, *236*(8), 2359–2371.
Littman, M. L. (2009). A tutorial on partially observable markov decision processes. *Journal of Mathematical Psychology*, *53*(3), 119–125.
Lopez, M., Balleine, B., & Dickinson, A. (1992). Incentive learning and the motivational control of instrumental performance by thirst. *Animal Learning & Behavior*, *20*(4), 322–328.
Maisto, D., Friston, K., & Pezzulo, G. (2019). Caching mechanisms for habit formation in active inference. *Neurocomputing*.
Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological science*, *24*(12), 2351–2360.
Martin, J. J. (1967). *Bayesian decision problems and markov chains*. Wiley.
Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A bayesian foundation for individual learning under uncertainty. *Frontiers in human neuroscience*, *5*.

Mazur, J. E. (2015). *Learning and behavior: Instructor's review copy*. Psychology Press.
McGeorge, A., & Faull, R. (1989). The organization of the projection from the cerebral cortex to the striatum in the rat. *Neuroscience*, *29*(3), 503–537.
McKim, T. H., Bauer, D. J., & Boettiger, C. A. (2016). Addiction history associates with the propensity to form habits. *Journal of cognitive neuroscience*, *28*(7), 1024–1038.
Meltzer, T., Yanover, C., & Weiss, Y. (2005). Globally optimal solutions for energy minimization in stereo vision using reweighted belief propagation. In *Computer vision, 2005. iccv 2005. tenth ieee international conference on* (Vol. 1, pp. 428–435).
Meyniel, F., Schlunegger, D., & Dehaene, S. (2015). The sense of confidence during probabilistic learning: A normative account. *PLoS computational biology*, *11*(6), e1004305.
Miller, K. J., Ludvig, E. A., Pezzulo, G., & Shenhav, A. (2018). Realigning models of habitual and goal-directed decision-making. In *Goal-directed decision making* (pp. 407–428). Elsevier.
Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without values. *Psychological review*.
Monahan, G. E. (1982). State of the art—a survey of partially observable markov decision processes: theory, models, and algorithms. *Management Science*, *28*(1), 1–16.
Montague, P. R., Hyman, S. E., & Cohen, J. D. (2004). Computational roles for dopamine in behavioural control. *Nature*, *431*(7010), 760.
Morris, R. W., Dezfouli, A., Griffiths, K. R., & Balleine, B. W. (2014). Action-value comparisons in the dorsolateral prefrontal cortex control choice between goal-directed actions. *Nature communications*, *5*(1), 1–10.
Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. The MIT Press.
Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, *30*(37), 12366–12378.
Neal, D. T., Wood, W., Labrecque, J. S., & Lally, P. (2012). How do habits guide behavior? perceived and actual triggers of habits in daily life. *Journal of Experimental Social Psychology*, *48*(2), 492–498.
Neal, R. M. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, *2*(11), 2.
Nebe, S., Kroemer, N. B., Schad, D. J., Bernhardt, N., Sebold, M., Müller, D. K., … Huys, Q. J. (2018). No association of goal-directed and habitual control with alcohol consumption in young adults. *Addiction biology*, *23*(1), 379–393.
Niv, Y., Joel, D., & Dayan, P. (2006). A normative perspective on motivation. *Trends in cognitive sciences*, *10*(8), 375–381.
O'doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *science*, *304*(5669), 452–454.
Ott, T., & Stoop, R. (2007). The neurodynamics of belief propagation on binary markov random fields. In *Advances in neural information processing systems* (pp. 1057–1064).
Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences*, *110*(52), 20941–20946.
Otto, A. R., Skatova, A., Madlon-Kay, S., & Daw, N. D. (2014). Cognitive control predicts use of model-based reinforcement learning. *Journal of cognitive neuroscience*, *27*(2), 319–333.
Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature communications*, *6*, 8096.
Parkes, S. L., Ravassard, P. M., Cerpa, J.-C., Wolff, M., Ferreira, G., & Coutureau, E. (2018). Insular and ventrolateral orbitofrontal cortices differentially contribute to goal-directed

behavior in rodents. *Cerebral Cortex*, *28*(7), 2313–2325.
Parkinson, J. A., Cardinal, R. N., & Everitt, B. J. (2000). Limbic cortical-ventral striatal systems underlying appetitive conditioning. In *Progress in brain research* (Vol. 126, pp. 263–285). Elsevier.
Payzan-LeNestour, E., Dunne, S., Bossaerts, P., & O'Doherty, J. P. (2013). The neural representation of unexpected uncertainty during value-based decision making. *Neuron*, *79*(1), 191–201.
Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.
Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier.
Pezzulo, G., Rigoli, F., & Chersi, F. (2013). The mixed instrumental controller: using value of information to combine habitual choice and mental simulation. *Frontiers in psychology*, *4*, 92.
Pollack, A. E. (2001). Anatomy, physiology, and pharmacology of the basal ganglia. *Neurologic clinics*, *19*(3), 523–534.
Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.
Rangel, A., & Hare, T. (2010). Neural computations associated with goal-directed choice. *Current Opinion in Neurobiology*, *20*(2), 262 - 270. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0959438810000371` (Cognitive neuroscience) doi: https://doi.org/10.1016/j.conb.2010.03.001
Redish, A. D., Jensen, S., Johnson, A., & Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychological review*, *114*(3), 784.
Reep, R. L., Cheatwood, J. L., & Corwin, J. V. (2003). The associative striatum: organization of cortical projections to the dorsocentral striatum in rats. *Journal of Comparative Neurology*, *467*(3), 271–292.
Renteria, R., Baltz, E. T., & Gremel, C. M. (2018). Chronic alcohol exposure disrupts top-down control over basal ganglia action selection to produce habits. *Nature communications*, *9*(1), 211.
Rescorla, R. A., & Wagner. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, *2*, 64–99.
Rummery, G. A., & Niranjan, M. (1994). *On-line q-learning using connectionist systems* (Vol. 37). University of Cambridge, Department of Engineering Cambridge, UK.
Rushworth, M. F., & Behrens, T. E. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature neuroscience*, *11*(4), 389.
Sage, J. R., & Knowlton, B. J. (2000). Effects of us devaluation on win–stay and win–shift radial maze performance in rats. *Behavioral neuroscience*, *114*(2), 295.
Schwartenbeck, P., FitzGerald, T. H., & Dolan, R. (2016). Neural signals encoding shifts in beliefs. *NeuroImage*, *125*, 578–586.
Schwartenbeck, P., FitzGerald, T. H., Mathys, C., Dolan, R., & Friston, K. (2014). The dopaminergic midbrain encodes the expected certainty about desired outcomes. *Cerebral Cortex*, *25*(10), 3434–3445.
Schwartenbeck, P., FitzGerald, T. H., Mathys, C., Dolan, R., Kronbichler, M., & Friston, K. (2015). Evidence for surprise minimization over value maximization in choice behavior. *Scientific reports*, *5*.

Schwöbel, S., Kiebel, S., & Marković, D. (2018). Active inference, belief propagation, and the bethe approximation. *Neural computation*, *30*(9), 2530–2567.

Schwöbel, S., Markovic, D., Smolka, M. N., & Kiebel, S. J. (2019). Balancing control: a bayesian interpretation of habitual and goal-directed behavior. *bioRxiv*, 836106.

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th python in science conference*.

Seger, C. A., & Spiering, B. J. (2011). A critical review of habit learning and the basal ganglia. *Frontiers in systems neuroscience*, *5*, 66.

Shon, A. P., & Rao, R. P. (2005). Implementing belief propagation in neural circuits. *Neurocomputing*, *65*, 393–399.

Simon, H. A. (1990). Invariants of human behavior. *Annual review of psychology*, *41*(1), 1–20.

Skinner, B. F. (1948). 'superstition'in the pigeon. *Journal of experimental psychology*, *38*(2), 168.

Smith, K. S., & Graybiel, A. M. (2013). A dual operator view of habitual behavior reflecting cortical and striatal dynamics. *Neuron*, *79*(2), 361–374.

Smith, K. S., & Graybiel, A. M. (2014). Investigating habits: strategies, technologies and models. *Frontiers in behavioral neuroscience*, *8*, 39.

Smith, K. S., & Graybiel, A. M. (2016). Habit formation. *Dialogues in clinical neuroscience*, *18*(1), 33.

Smith, K. S., Virkud, A., Deisseroth, K., & Graybiel, A. M. (2012). Reversible online control of habitual behavior by optogenetic perturbation of medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, *109*(46), 18932–18937.

Smittenaar, P., FitzGerald, T. H., Romei, V., Wright, N. D., & Dolan, R. J. (2013). Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron*, *80*(4), 914–919.

Solway, A., & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. *Psychological review*, *119*(1), 120.

Steimer, A., Maass, W., & Douglas, R. (2009). Belief propagation in networks of spiking neurons. *Neural Computation*, *21*(9), 2502–2523.

Sudderth, E. B., Mandel, M. I., Freeman, W. T., & Willsky, A. S. (2004). Visual hand tracking using nonparametric belief propagation. In *Computer vision and pattern recognition workshop, 2004. cvprw'04. conference on* (pp. 189–189).

Sutton, R. S., & Barto, A. G. (1998a). *Introduction to reinforcement learning* (Vol. 135). MIT press Cambridge.

Sutton, R. S., & Barto, A. G. (1998b). *Reinforcement learning: An introduction* (Vol. 1). MIT press Cambridge.

Talmi, D., Seymour, B., Dayan, P., & Dolan, R. J. (2008). Human pavlovian–instrumental transfer. *Journal of Neuroscience*, *28*(2), 360–368.

Thorn, C. A., Atallah, H., Howe, M., & Graybiel, A. M. (2010). Differential dynamics of activity changes in dorsolateral and dorsomedial striatal loops during learning. *Neuron*, *66*(5), 781–795.

Thorndike, E. L. (1898). Animal intelligence: an experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, *2*(4), i.

Thrailkill, E. A., & Bouton, M. E. (2015). Contextual control of instrumental actions and habits. *Journal of Experimental Psychology: Animal Learning and Cognition*, *41*(1), 69.

Todorov, E. (2009). Efficient computation of optimal actions. *Proceedings of the national academy of sciences*, *106*(28), 11478–11483.

Tricomi, E., Balleine, B. W., & O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience*, *29*(11), 2225–2232.

Tricomi, E. M., Delgado, M. R., & Fiez, J. A. (2004). Modulation of caudate activity by action contingency. *Neuron*, *41*(2), 281–292.

Tsutsui, K.-I., Oyama, K., Nakamura, S., & Iijima, T. (2016). Comparative overview of visuospatial working memory in monkeys and rats. *Frontiers in systems neuroscience*, *10*, 99.

The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. (2000). *Cognitive Psychology*, *41*(1), 49 - 100. Retrieved from `http://www.sciencedirect.com/science/article/pii/S001002859990734X` doi: https://doi.org/10.1006/cogp.1999.0734

Valentin, V. V., Dickinson, A., & O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *Journal of Neuroscience*, *27*(15), 4019–4026.

Verplanken, B., & Roy, D. (2016). Empowering interventions to promote sustainable lifestyles: Testing the habit discontinuity hypothesis in a field experiment. *Journal of Environmental Psychology*, *45*, 127–134.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … Contributors, S. . . (2019, Jul). SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. *arXiv e-prints*, arXiv:1907.10121.

Volkow, N. D., & Morales, M. (2015). The brain on drugs: from reward to addiction. *Cell*, *162*(4), 712–725.

Vossel, S., Mathys, C., Daunizeau, J., Bauer, M., Driver, J., Friston, K. J., & Stephan, K. E. (2013). Spatial attention, precision, and bayesian inference: a study of saccadic response speed. *Cerebral cortex*, *24*(6), 1436–1450.

Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, *1*(1–2), 1–305.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards* (Unpublished doctoral dissertation). King's College, Cambridge.

Watson, P., & de Wit, S. (2018). Current limits of experimental research into habits and future directions. *Current opinion in behavioral sciences*, *20*, 33–39.

Watson, P., Wiers, R., Hommel, B., & De Wit, S. (2014). Working for food you don't desire. cues interfere with goal-directed food-seeking. *Appetite*, *79*, 139–148.

Weiss, Y. (2001). Comparing the mean field method and belief propagation for approximate inference in mrfs. *Advanced mean field methods: theory and practice*, 229–240.

Wilson, R. C., Takahashi, Y. K., Schoenbaum, G., & Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron*, *81*(2), 267–279.

Wood, W., & Rünger, D. (2016). Psychology of habit. *Annual review of psychology*, *67*, 289–314.

Yedidia, J. S., Freeman, W. T., & Weiss, Y. (2001a). Bethe free energy, kikuchi approximations, and belief propagation algorithms. *Advances in neural information processing systems*, *13*.

Yedidia, J. S., Freeman, W. T., & Weiss, Y. (2001b). Generalized belief propagation. In *Advances in neural information processing systems* (pp. 689–695).

Yedidia, J. S., Freeman, W. T., & Weiss, Y. (2003a). Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, *8*, 236–239.

Yedidia, J. S., Freeman, W. T., & Weiss, Y. (2003b). Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, *8*, 236–239.

Yedidia, J. S., Freeman, W. T., & Weiss, Y. (2005). Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on information theory*, *51*(7), 2282–2312.

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, *7*(6), 464.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European journal of neuroscience*, *19*(1), 181–189.

Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience*, *22*(2), 513–523.

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(4), 681–692.

Yu, S.-Z., & Kobayashi, H. (2003). An efficient forward-backward algorithm for an explicit-duration hidden markov model. *IEEE signal processing letters*, *10*(1), 11–14.

Yuille, A., & Kersten, D. (2006). Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, *10*(7), 301–308.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: a mind-brain perspective. *Psychological bulletin*, *133*(2), 273.

Declaration

I hereby certify that I have authored this Dissertation entitled “Bayesian cognitive modeling of the balancing between goal-directed and habitual behavior” independently and without undue assistance from third parties. No other than the resources and references indicated in this thesis have been used. I have marked both literal and accordingly adopted quotations as such. There were no additional persons involved in the intellectual preparation of the present thesis. I am aware that violations of this declaration may lead to subsequent withdrawal of the degree.
I furthermore herewith declare that I recognize the Regulations for Obtaining a Doctoral Degree of the TU Dresden department of mathematics and natural sciences.

Dresden, 16th June 2020

Sarah Schwöbel